CONCENTRATION INEQUALITIES FOR DEPENDENT RANDOM VARIABLES

Daniel Paulin

(M.Sc., ECP Paris; B.Sc., BUTE Budapest)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICS

NATIONAL UNIVERSITY OF SINGAPORE

2014

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in this thesis.

This thesis has also not been submitted for any degree in any university previously.

Daniel Paulin

December 2, 2014

Acknowledgements

First and foremost, I would like to thank my advisors, Louis Chen and Adrian Röllin, for the opportunity to study in Singapore, and their guidance during my thesis. I am deeply indebted to them for all the discussions, which have helped me to progress in my research and improved my presentation and writing skills. I am also grateful to Professor Chen for making possible for me to participate in the ICM 2010 in India, and the workshop "Concentration Inequalities and their Applications" in France.

During my years at NUS, my advisors and colleagues have organised several working seminars on various topics. These have been very helpful, and I would like to thank some of the speakers, Sun Rongfeng, Fang Xiao, Sanjay Chaudhuri, Siva Athreya, Ajay Jasra, Alexandre Thiery, Alexandros Beskos, and David Nott.

I am indebted to all my collaborators and colleagues for the discussions. Special thanks go to Benjamin Gyori, Joel A. Tropp, and Lester Mackey. After making some of my work publicly available, I have received valuable feedback and encouragement from several people. I am particularly grateful to Larry Goldstein, Daniel Rudolf, Yann Ollivier, Katalin Márton, Malwina Luczak, and Laurent Saloff-Coste.

I am greatly indebted to my university teachers in Hungary, in particular, Domokos Szász and Mogyi Tóth, for infecting me with their enthusiasm of probability, and to Péter Moson, for his help with my studies in France. I am also greatly indebted to

my high school teachers from the wonderful Fazekas Mihály Secondary School, especially to Tünde Fazakas, András Hraskó, László Surányi, and Gábor Horváth. I thank Sándor Róka, a good friend of my family, for his wonderful books.

An outstanding math teacher who had a great influence on my life is Lajos Pósa, the favourite student of Paul Erdős. Thank you very much for your support all these years!

My PhD years have been made colourful by my friends and flatmates in Singapore. Thank you Alexandre, Susan, Benjamin, Claire, Andras, Aggie, Brad, Rea, Jeroen, Max, Daikai, and Yvan for the great environment.

I have infinite gratitude towards my parents for bringing me up, and for their constant encouragement and support, and I am very grateful to my brother Roland for our discussions. Finally, this thesis would have never been written without the love of my wife candidate, Dandan.

To my family.

Contents

Summary

This thesis contains contributions to the theory of concentration inequalities, in particular, concentration inequalities for dependent random variables. In addition, a new concept of spectral gap for non-reversible Markov chains, called pseudo spectral gap, is introduced.

We consider Markov chains, stationary distributions of Markov chains (including the case of dependent random variables satisfying the Dobrushin condition), and locally dependent random variables. In each of these cases, we prove new concentration inequalities that improve considerably those in the literature. In the case of Markov chains, we prove concentration inequalities that are only the mixing time of the chain times weaker than those for independent random variables. In the case of stationary distributions of Markov chains, we show that Lipschitz functions are highly concentrated for distributions arising from fast mixing chains, if the chain has small step sizes. For locally dependent random variables, we prove concentration inequalities under several different types of local dependence.

List of Figures

List of Symbols

The following description explains the meaning of the most frequently used symbols in this thesis. Note that there are a few places where some of these symbols have a slightly different usage.

Symbol	**Description**
$\mathbb{R}^k$	k dimensional Euclidean space
$\mathbb{R}_+$	set of positive real numbers
$\mathbb{C}$	set of complex numbers
$\mathbb{Z}$	set of integers
$\mathbb{N}$	set of natural numbers
X	a random vector, with coordinates $X = (X_1, \ldots, X_n)$
Λ	state space of a random vector, of the form $\Lambda = \Lambda_1 \times \ldots \times \Lambda_n$
Ω	state space of a random vector, of the form $\Omega = \Omega_1 \times \ldots \times \Omega_n$
$\mathbb{P}$	probability distribution induced by the random vector X
$\mathbb{E}$	expected value
$\mathcal{L}(X\|Y = y)$	law of a random vector X conditioned on the event that the random vector Y takes value y
$d_{\mathrm{TV}}(\mu, \nu)$	total variational distance of two probability distributions μ and ν

$P(x, \mathrm{d}y)$	a Markov kernel
π	stationary distribution of a Markov kernel
$L^k(\pi)$	set of measurable functions f such that $\lvert f\rvert^k$ is integrable with respect to the distribution π
L^k	set of measurable functions f on $\mathbb{R}^n$ such that $\lvert f\rvert^k$ is integrable with respect to the Lebesgue measure on $\mathbb{R}^n$
t_{mix}	mixing time of a Markov chain
γ	spectral gap of a Markov chain
γ_{ps}	pseudo spectral gap of a Markov chain
$\langle a, b\rangle$	scalar product of two vectors
$\langle f, g\rangle_\pi$	scalar product for $f, g \in L^2(\pi)$, $\langle f, g\rangle_\pi := \int_x f(x)g(x)\pi(\mathrm{d}x)$.
$\lVert A\rVert_k$	L^k norm of the matrix A
$\lVert A\rVert_{2,\pi}$	operator norm of A as an operator on $L^2(\pi)$
$\{X(k)\}_{k=0,1,\ldots}$	a realisation of a Λ valued Markov chain
$X_i(k)$	ith coordinate of the random vector $X(k)$
κ	coarse Ricci curvature
κ_k	multi-step coarse Ricci curvature

Chapter 1

Introduction

Concentration inequalities are bounds on the quantity $\mathbb{P}(f(X) - \mathbb{E}(f(X)) \geq t)$, where X is typically a vector of random variables $X := (X_1, \ldots, X_n)$. The case where X is a vector of independent random variables is well-understood, and many inequalities are rather sharp in this case (see the introductionary book by Boucheron, Lugosi, and Massart (2013b)). Applications of such inequalities are numerous and can be found in computer science, statistics, and probability theory.

In stark contrast, in the case of dependent random variables, the results in the literature are often not sharp, even for some of the most frequently occurring types of dependence. Because of this, there seem to be much fewer applications of such inequalities as compared to the independent case.

In this thesis, we sharpen and extend such inequalities for some important dependency structures, namely Markov chains, stationary distributions of Markov chains, and local dependence.

A classical example of concentration inequalities is McDiarmid's bounded differences inequality. Let Ω be a Polish space, let $X = (X_1, \ldots, X_n)$ be a vector of

independent random variables taking values in Ω^n, and let $f : \Omega^n \to \mathbb{R}$ be a function such that changing the value of coordinate i can change the value of f at most by c_i, for $1 \leq i \leq n$. Then

$$\mathbb{P}(|f(X) - \mathbb{E}(f)| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right), \tag{1.0.1}$$

where $\mathbb{E}(f) := \mathbb{E}(f(X))$. The importance of this result lies in the fact that, whereas the range of f satisfies that $\sup_{x\in\Omega^n} f(x) - \inf_{x\in\Omega^n} f(x) \leq \sum_{i=1}^n c_i$, the typical size of the deviation $|f(X) - \mathbb{E}(f)|$ is only $(\sum_{i=1}^n c_i^2)^{1/2}$, which can be much smaller. Thus the bound expresses the fact that if f is a function that depends only a "little bit" on each of its coordinates and n is large, then $f(X)$ is concentrated around its mean at a much smaller range than its maximal possible deviation.

Inequality (1.0.1) implies, in particular, Hoeffding's inequality. Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables with expectation $\mathbb{E}(X_1)$, satisfying $a \leq X_i \leq b$ almost surely. Hoeffding's inequality states that for any $t \geq 0$,

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n} - \mathbb{E}(X_1)\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2 n}{(b-a)^2}\right). \tag{1.0.2}$$

This can be obtained from the (1.0.1) by considering the function $f(x) = (x_1 + \ldots + x_n)/n$.

A similar inequality, that also taking into account the variances of X_i, is Bernstein's inequality. Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables, with expectation $\mathbb{E}(X_1)$, satisfying $|X_i - \mathbb{E}(X_i)| \leq C$ almost surely, then for any $t \geq 0$,

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n X_i}{n} - \mathbb{E}(X_1)\right| \geq t\right) \leq 2\exp\left(-\frac{t^2 n}{2\mathrm{Var}(X_1) + (2/3)Ct}\right). \tag{1.0.3}$$

Then typically this is sharper than (1.0.2), especially when $\mathrm{Var}(X_1) \ll C^2$.

Hoeffding's and Bernstein's inequalities are useful for constructing non-asymptotically valid confidence intervals of $\mathbb{E}(X_1)$, given n independent samples $X_1, \ldots, X_n$, by comparing the difference between the estimated mean $\hat{X} = (\sum_{i=1}^n X_i)/n$ and the mean $\mathbb{E}(X_1)$. In the particular case of Bernoulli random variables with parameter p, $\mathbb{E}(X_1) = p$, and Hoeffding's inequality states that $\mathbb{P}(|\hat{X} - p| \geq t) \leq 2\exp(-2t^2 \cdot n)$. This means that the typical deviations are of order $\sqrt{n}$.

In many practical situations, however, independent sampling is not possible, and the only way to sample from the distribution of interest is via the Markov Chain Monte Carlo method, in which case $X_1, \ldots, X_n$ is a realisation of a Markov chain. Suppose that a Markov chain takes values in a Polish state space Ω, has unique stationary distribution π, and that we are interested in evaluating the expectation of some function $f : \Omega \to \mathbb{R}$. Then we can use the approximation $(\sum_{i=1}^n f(X_i))/n \approx \mathbb{E}_\pi(f)$ to evaluate the expectation. Now it is of great practical importance to know how good is this approximation, since this determines how many samples do we need from the Markov chain, and hence how long do we need to run our simulation. For this reason, it is important to generalise the concentration inequalities above to the case where $X_1, \ldots, X_n$ is a Markov chain.

It seems that, unlike in the independent case, where many of the sharp results known can be obtained by log-Sobolev inequalities and the entropy method, different types of dependences and different types of functions require different methods to get sharp bounds.

In order to get sharp concentration bounds for Markov chains, we need to understand their mixing properties. One way to express the mixing properties of Markov

chains is by analysing their spectrum. Let $L^2(\pi)$ be the Hilbert space of measurable functions $f : \Omega^n \to \mathbb{R}$ that are square integrable with respect to π, equipped with the scalar product $\langle f, g \rangle_\pi = \mathbb{E}_\pi(fg)$. Then the Markov kernel $\boldsymbol{P}$ defined as $\boldsymbol{P}(f)(x) = \mathbb{E}(f(X_2)|X_1 = x)$ is a linear operator on this space. In the case of reversible chains, this operator is self-adjoint, and thus its eigenvalues are real. As it is well known, the Markov kernel's largest eigenvalue is always one. The *spectral gap*, denoted by $\gamma = \gamma(\boldsymbol{P})$, is essentially the distance between its largest and second largest eigenvalue. We denote by γ^* the *absolute spectral gap* of the chain, which is essentially the gap between 1 and the eigenvalue with the second largest absolute value.

In the case of non-reversible chains, the eigenvalues of $\boldsymbol{P}$ may be complex. The standard approach in the literature in this case is to look at the spectral gap of the multiplicative reversiblication $\boldsymbol{P}^*\boldsymbol{P}$, denoted by $\gamma(\boldsymbol{P}^*\boldsymbol{P})$ (here $\boldsymbol{P}^*$ denotes the adjoint of $\boldsymbol{P}$, defined by the Markov kernel $P^*(x, dy) := \frac{P(y,dx)}{\pi(dx)} \cdot \pi(dy)$). This corresponds to the spectral gap of the Markov chain created from the original chain by taking one step forward in time, followed by one step backward in time.

Another way to express mixing properties of Markov chains is by means of mixing times. The total variational distance mixing time, denoted by t_{mix} is the most frequently used in the literature. It equals to the number of steps the chain has to take to get to less that $1/4$ in total variational distance to the stationary distribution from any initial point.

For reversible chains, the mixing time and the spectral gap are related by some simple inequalities, stating that whenever the mixing time is small, the spectral gap is large, and in the case of chains with finite state spaces, that whenever the spectral

gap is large, the mixing time is small (we will discuss this in more details in Chapter 3). In practice, $1/\gamma$ and t_{mix} are typically of the same orders of magnitude up to logarithmic factors.

For non-reversible chains on finite state spaces, it is also known that whenever $\gamma(P^*P)$ is large, t_{mix} is small. However, the converse is not true, since there are chains that mix fast in total variational distance (i.e. t_{mix} is small), but for which $\gamma(\boldsymbol{P}^*\boldsymbol{P}) = 0$. This has lead us to propose a new definition of spectral gap for non-reversible chains. Let the *pseudo spectral gap* of the chain be defined as

$$\gamma_{\text{ps}} := \max_{k \geq 0} \gamma\left((\boldsymbol{P}^*)^k \boldsymbol{P}^k\right)/k.$$

We are going to show that this quantity behaves similarly to the spectral gap for reversible chains. That is, if the mixing time is small, the pseudo spectral gap is large, and for chains on finite state spaces, if the pseudo spectral gap is large, the mixing time is small.

In Chapter 3, we prove concentration inequalities for functions of Markov chains. We use two different methods to prove these inequalities for sums, and more general functions. In the case of general functions, we use what we call *Marton couplings*, originally introduced by Marton (2003). Using this coupling, and by partitioning the random variables into larger blocks of size proportional to the mixing time, we generalise the martingale-type approach of Chazottes, Collet, Külske, and Redig (2007). This leads to the following generalisation of McDiarmid's bounded differences inequality to Markov chains, with constants that are proportional to the mixing time of the chain. If $X = (X_1, \ldots, X_n)$ is a Markov chain on the state space Ω, and $f : \Omega^n \to \mathbb{R}$ is a function such that changing the value of coordinate i can change the value of f

at most by c_i, for $1 \leq i \leq n$, then for any $t \geq 0$,

$$\mathbb{P}(|f(X) - \mathbb{E}(f)| \geq t) \leq 2\exp\left(-\frac{t^2}{4.5t_{\text{mix}} \cdot \sum_{i=1}^n c_i^2}\right). \tag{1.0.4}$$

The Central Limit Theorem implies that under mild conditions, $(\sum_{i=1}^n f(X_i))/\sqrt{n}$ converges in distribution to $N(\mathbb{E}_\pi(f), \sigma_{\text{as}}^2)$, where σ_{as}^2 denotes the asymptotic variance of a function $f : \Omega \to \mathbb{R}$, defined as

$$\sigma_{\text{as}}^2 := \lim_{n\to\infty} \frac{1}{n}\text{Var}\left(\sum_{i=1}^n f(X_i)\right).$$

We propose a new estimator to this quantity (based on $f(X_1), \ldots, f(X_n)$). Our estimator is a rather complicated function of $X_1, \ldots, X_n$, however, we show that it satisfies the conditions of our version of McDiarmid's bounded differences inequality, and deduce that it is highly concentrated. This allows us to estimate σ_{as}^2 with arbitrary precision by setting n sufficiently high (depending on the mixing time of the chain).

Using spectral methods due to Lezaud (1998b), we obtain concentration bounds for sums of the form $\sum_{i=1}^n f(X_i)$, and more generally, of form $\sum_{i=1}^n f_i(X_i)$, for a Markov chain $X_1, \ldots, X_n$. We obtain that for a stationary and reversible Markov chain with spectral gap γ, and a function f satisfying $|f(x) - \mathbb{E}(f)| \leq C$ for some constant $C > 0$,

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n f(X_i)}{n} - \mathbb{E}(f)\right| \geq t\right) \leq 2\exp\left(-\frac{t^2 \cdot n}{2\left(\sigma_{\text{as}}^2 + 0.8\text{Var}(f)\right) + 10(C/\gamma) \cdot t}\right). \tag{1.0.5}$$

This is a type of Bernstein inequality. For small values of t, this bound is roughly equal to $\exp\left(-\frac{t^2 \cdot n}{2(\sigma_{\text{as}}^2 + 0.8\text{Var}(f))}\right)$. For a standard normal random variable, the sharpest

tail bound that holds is of the form $\exp(-t^2/2)$. Since the Central Limit Theorem implies that $(\sum_{i=1}^n f(X_i))/n$ is close to $N(\mathbb{E}_\pi(f), \sigma_{\text{as}}^2/n)$ in distribution, the sharpest tail bound that we can expect is of the form $\exp\left(-\frac{t^2 \cdot n}{2\sigma_{\text{as}}^2}\right)$. Therefore our bound is essentially sharp for small values of t (except for the $0.8\text{Var}(f)$ term, but typically this is much smaller than σ_{as}^2). The Bernstein inequality of Lezaud (1998b) for reversible chains only depends on γ and $\text{Var}(f)$, but does not incorporates the asymptotic variance σ_{as}, meaning that our bound is sharper.

For stationary non-reversible chains, using the pseudo spectral gap, we obtain the following version of Bernstein's inequality. Under the same conditions as in (1.0.5), for any $t \geq 0$,

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n f(X_i)}{n} - \mathbb{E}(f)\right| \geq t\right) \leq \exp\left(-\frac{t^2 \cdot \gamma_{\text{ps}} \cdot (n - 1/\gamma_{\text{ps}})}{8\text{Var}(f) + 20Ct}\right). \tag{1.0.6}$$

The Bernstein inequality of Lezaud (1998b) uses the spectral gap of the multiplicative reversiblication, $\gamma(\boldsymbol{P}^*\boldsymbol{P})$, thus our bound is sharper.

The main application of the bounds (1.0.5) and (1.0.6) is to estimate the error of MCMC empirical averages (that is the quality of the approximation $\mathbb{E}_\pi(f) \approx \frac{1}{n}\sum_{i=1}^n f(X_i)$).

We include generalisations of McDiarmid and Bernstein-type concentration inequalities to Markov processes. The proofs for this case are based on simple limiting arguments.

In addition to Markov chains, there are other dependency structures that can arise in practice, and are thus worth studying. One insightful way of looking at distributions of dependent random variables is by considering a Markov chain that has this distribution as its stationary distribution. There are several approaches in the

literature that show that under various conditions on the mixing rate of this Markov chain (so-called *contraction conditions*), the stationary distribution satisfies concentration inequalities (see Chatterjee (2005), Ollivier (2009), and Djellout, Guillin, and Wu (2004)). In Chapter 4, we generalise Ollivier's coarse Ricci curvature approach, and also identify connections to the results of Chatterjee (2005).

Let us consider a stationary Markov chain with transition kernel P on a Polish space Ω equipped with a metric $d : \Omega^2 \to \mathbb{R}$ (which we denote by (Ω, d)), with stationary distribution π. Denote the distribution of one step in the Markov chain starting from $x \in \Omega$ by P_x. Given two measures μ and ν on Ω, we define their Wasserstein distance $W_1(\mu, \nu)$ as

$$W_1(\mu, \nu) := \inf_{\xi \in \Pi(\mu,\nu)} \int_{(x,y)\in\Omega^2} d(x, y) \mathrm{d}\xi(x, y),$$

with $\Pi(\mu, \nu)$ denoting the set of distributions on Ω^2 with marginals μ and ν.

A natural way to quantify the mixing rate is to compare $W_1(P_x, P_y)$ with $d(x, y)$. Following Ollivier (2009), we define the *coarse Ricci curvature* κ to be the largest possible constant such for any two disjoint $x, y \in \Omega$, $W_1(P_x, P_y) \leq (1 - \kappa)d(x, y)$ (it is easy to see that this constant always exists, but may be $-\infty$). If $\kappa > 0$, then it can be thought as a kind of *contraction coefficient*, since after k steps in the chain, we have $W_1(P_x^k, P_y^k) \leq (1 - \kappa)^k$. Here P_x^k denotes the distribution of the kth step of the Markov chain starting from x.

This property is then used to prove concentration for Lipschitz functions. Ollivier (2009) shows that under the assumption $\kappa > 0$, for $X \sim \pi$ and for some range

$0 \leq t \leq t_{\max}$, they satisfy concentration inequalities of the form

$$\mathbb{P}(f(X) - \mathbb{E}(f) \geq t) \leq \exp\left(-\frac{t^2 \cdot n}{6\sigma^2 \cdot (1/\kappa) \cdot \|f\|_{\text{Lip}}^2}\right), \tag{1.0.7}$$

where σ^2 is a quantity related to the typical size of the jumps of the Markov chain, n is a quantity related to the dimension of the space, and $\|f\|_{\text{Lip}}$ is the Lipschitz coefficient of f.

In this thesis, we generalise this bound by considering the coarse Ricci curvature of multiple steps in the Markov chain. Define P_x^k as the distribution of taking k steps in the chain, starting from x, and let the *multi-step coarse Ricci curvature* κ_k be the largest real number such that $W_1(P_x^k, P_y^k) \leq (1 - \kappa_k)d(x, y)$. Then we show that concentration inequalities of the type

$$\mathbb{P}(f(X) - \mathbb{E}(f) \geq t) \leq \exp\left(-\frac{t^2 \cdot n}{6\sigma^2 \cdot \kappa_\Sigma^{(2)} \cdot \|f\|_{\text{Lip}}^2}\right), \tag{1.0.8}$$

hold for some range $0 \leq t \leq t_{\max}$, with $\kappa_\Sigma^{(2)} := \sum_{k=0}^{\infty}(1 - \kappa_k)^2$. It is easy to see that for $\kappa > 0$, $\kappa_\Sigma^{(2)} < 1/\kappa$, implying that our result is stronger then (1.0.7). We are going to give examples for $\kappa > 0$, but where $\kappa_\Sigma^{(2)}$ is much smaller than $1/\kappa$, and examples where $\kappa < 0$, but where $\kappa_\Sigma^{(2)}$ is finite.

The coarse Ricci curvature has connections with the spectral properties of Markov chains. For reversible chains it is known that $\gamma \geq \kappa$. Here we generalise this result and show that $\gamma \geq \kappa_k/k$, and also show how to bound the pseudo spectral gap γ_{ps} in terms of the coarse Ricci curvature κ_k.

We include applications to the split-merge walk on random partitions, Glauber dynamics on statistical physical spin models, and a random walk on the binary cube

with a forbidden region.

Although the multi-step coarse Ricci curvature approach works for many dependency structures, one of its disadvantages is that the concentration bounds only take into account the Lipschitz coefficient of f. For more complicated functions, Talagrand's convex distance inequality can yield better bounds. In Chapter 5, we will prove a version of Talagrand's convex distance inequality for weakly dependent random variables satisfying the so-called Dobrushin condition. We show that, in particular, sampling without replacement satisfies this condition. Our approach is an extension of the method of Chatterjee (2005), which is based on Stein's method of exchangeable pairs. We give applications to classical problems from computer science, the stochastic travelling salesman problem, and the Steiner tree problem.

In Chapter 5, similarly to Chatterjee (2005), we use exchangeable pairs to prove concentration inequalities. Chen and Röllin (2010) has introduced a more general coupling structure called Stein coupling, defined as follows.

Definition 1.0.1. Let (W, W', G) be a coupling of square integrable random variables. We call (W, W', G) a Stein coupling if

$$\mathbb{E}\{Gf(W') - Gf(W)\} = \mathbb{E}\{Wf(W)\},$$

for all functions for which the expectation exists.

Exchangeable pairs are a special case of this coupling structure. From the definition, it is easy to show that the moment generating function $m(\theta) = \mathbb{E}(e^{\theta W})$ satisfies

$$m'(\theta) = \mathbb{E}\left\{G\left(e^{\theta W'} - e^{\theta W}\right)\right\}, \tag{1.0.9}$$

which means that concentration inequalities can be obtained in terms of the typical size of G and $W - W'$. In Chapter 6, we show that non-exchangeable Stein couplings can also be used to prove concentration inequalities. We apply our results to random graph models, in particular, to the number of edges in geometric random graphs, and to randomly chosen large subgraphs of huge graphs.

Finally, in Chapter 7, we investigate concentration inequalities for locally dependent random variables. Let $[n] := \{1, \ldots, n\}$. We say that family of random variables $\{X_i\}_{1 \leq i \leq n}$ satisfies (LD) if for each $1 \leq i \leq n$ there exists $A_i \in [n]$ (called the neighbourhood of X_i) such that X_i and $\{X_j\}_{j \in A_i^c}$ are independent. We define the *dependency graph* of $\{X_i\}_{1 \leq i \leq n}$ as a graph with $[n]$ where i and j are interconnected if $i \in A_j$ or $j \in A_i$ (that is, X_i or X_j is in the neighborhood of the other).

(Janson, 2004) obtains concentration results for sums of random variables satisfying (LD), and also obtain Hoeffding and Bernstein inequalities, with constants that are only by the chromatic number of $\mathcal{G}$ times weaker than in the independent case. We show that unlike in the case of Hoeffding and Bernstein inequalities, (LD) dependence is not sufficient to show McDiarmid's bounded differences inequality. We define a stronger condition of local dependence, called (HD) dependence, and show that it does imply a version of the bounded differences inequality.

Now we are going to explain the organisation of this thesis. In Chapter 2, we introduce the subject of concentration inequalities, give some illustrative examples, and review the most popular methods for proving such inequalities. Chapter 3 contains our results for functions of Markov chains, which we obtain using Marton couplings, and spectral methods. Chapter 4 proves concentration inequalities for Lipschitz functions, when the measure arises as the stationary distribution of a fast-mixing Markov

chain. In Chapter 5, we will prove Talagrand's convex distance inequality for weakly dependent random variables satisfying the Dobrushin condition. Chapter 6 proves concentration inequalities based on Stein couplings. Finally, in Chapter 7 we will prove concentration inequalities for functions of locally dependent random variables.

Chapter 2

Review of the literature

In this chapter, we briefly review the literature of concentration inequalities. First, we explain the relation of the set formulation and the functional formulation of the concentration of measure phenomenon. After this, we start with a section containing selected examples of concentration inequalities, in particular, Hoeffding and Bernstein inequalities, with an application of Hoeffding's inequality to the running time of the Quicksort algorithm, followed by McDiarmid's bounded differences inequality, with an application to the chromatic number of the Erdős-Rényi random graph, then Talagrand's convex distance inequality, with an application to the concentration of the eigenvalues of random symmetric matrices, and finally the Gromov-Lévy inequality for concentration on a sphere. This is followed by a section about some of the most popular methods for proving concentration inequalities.

2.1 Concentration of sets versus functions

The first concentration inequalities were introduced by Bernstein (1924), Chernoff (1952), and later generalised by Hoeffding (1963) and Azuma (1967). The set formulation of the concentration of measure phenomenon was introduced by Milman in the early seventies, in the asymptotic theory of Banach spaces. Since then, it has found numerous applications in diverse fields such as geometry, functional analysis, discrete mathematics, and probability theory.

The standard reference on concentration inequalities is Ledoux (2001). Boucheron, Lugosi, and Massart (2013b) and Dubhashi and Panconesi (2009) are written at a more elementary level, and they contain many applications and exercises.

We illustrate the concentration of measure phenomenon with the example of concentration on a hypercube. Let $\Lambda := \{0,1\}^n$ be equipped with the counting measure μ, i.e. for any $A \subset \Lambda$, $\mu(A) := |A|/2^n$, where $|A|$ denotes the number of elements in A. For $x, y \in \Lambda$, $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n)$, let $d(x,y) := \sum_{i=1}^n \mathbb{1}[x_i \neq y_i]$ be the Hamming distance between x and y. For two sets $A, B \subset \Lambda$, we define the set distance $d(A,B) := \inf_{x \in A, y \in B} d(x,y)$ and let $d(x,B) := d(\{x\}, B)$. Then for any $A, B \subset \Lambda$,

$$\mu(A) \cdot \mu(B) \leq \exp\left(-\frac{d(A,B)^2}{2n}\right). \tag{2.1.1}$$

This inequality is the set formulation of the concentration of measure phenomenon. It says that if two sets are far from each other, then at least one of them has small probability.

Alternatively, suppose that $X = (X_1, \ldots, X_n)$ is a vector of i.i.d. Bernoulli random variables with parameter $1/2$. Denote the measure induced by X by $\mathbb{P}$. Suppose that

a function $f : \Lambda \to \mathbb{R}$ is 1-Lipschitz with respect to the Hamming distance d. Then for any $t \geq 0$,

$$\mathbb{P}(|f(X) - \mathbb{E}(f)| \geq t) \leq 2\exp(-2t^2/n). \tag{2.1.2}$$

Remark 2.1.1. More precisely, we have $\mathbb{P}(f(X) - \mathbb{E}(f) \geq t) \leq \exp(-2t^2/n)$ and $\mathbb{P}(f(X) - \mathbb{E}(f) \leq -t) \leq \exp(-2t^2/n)$. To avoid unnecessary repetition, the convention in the literature is to state the results in the form (2.1.2). Here we will adapt this convention.

This bound means that the typical deviation of the function f around its mean is $\sqrt{n}$ (meanwhile, the maximal deviation can be up to n). Such inequalities are the called the functional formulation of the concentration of measure phenomenon.

The two formulations are equivalent, up to small constant factors. Here we show this in the case of Gaussian tails. Note that Gaussian concentration (i.e. bounds of the form $\exp(-t^2/C)$) of f around its mean is equivalent to concentration around its median, as shown in Proposition 1.8. of Ledoux (2001).

Firstly, suppose that Λ is a Polish space equipped with a metric d, and $\mathbb{P}$ is a probability distribution on Λ such that for any two sets $A, B \in \Lambda$,

$$\mathbb{P}(A) \cdot \mathbb{P}(B) \leq \exp(-d(A, B)^2/C)$$

for some positive constant C. Let $X \sim \mathbb{P}$ be a Λ valued random variable. Suppose that $f : \Lambda \to \mathbb{R}$ is 1-Lipschitz with respect to d. Denote its median by $\mathbb{M}(f)$ (by this we mean any real number satisfying that $\mathbb{P}(f(X) \geq \mathbb{M}(f)) \geq 1/2$ and $\mathbb{P}(f(X) \leq \mathbb{M}(f)) \geq 1/2$). Let $A := \{x \in \Lambda : f(x) \leq \mathbb{M}(f)\}$, and for every $t > 0$, let $B_t := \{x \in \Lambda : f(x) \geq \mathbb{M}(f) + t\}$. Then by the 1-Lipschitz property of f, we

have $d(A, B_t) \geq t$, thus by our initial assumption, we obtain that $\mathbb{P}(A) \cdot \mathbb{P}(B_t) \leq \exp(-t^2/C)$. Now $\mathbb{P}(A) \geq 1/2$, thus we obtain that

$$\mathbb{P}(f(X) - \mathbb{M}(f) \geq t) \leq 2\exp(-t^2/C),$$

and the same bound holds for the lower tail too.

Alternatively, suppose that Lipschitz functions are concentrated around their median, i.e. $\mathbb{P}(f(X) - \mathbb{M}(f) \geq t) \leq 2\exp(-t^2/C)$ for every 1-Lipschitz f. Let A, B be two sets in Λ.

Suppose first that A has probability larger than $1/2$. Then the median of the 1-Lipschitz function $d(x, A)$ is 0, thus by our assumption,

$$\begin{aligned}\mathbb{P}(B) &\leq \mathbb{P}(d(x, A) \geq d(A, B)) = \mathbb{P}(d(x, A) \geq \mathbb{M}(d(x, A)) + d(A, B)) \\ &\leq 2\exp(-d(A, B)^2/C).\end{aligned}$$

Therefore $\mathbb{P}(A)\mathbb{P}(B) \leq 2\exp(-d(A, B)^2/C)$ in this case. The case when B has probability larger than $1/2$ is similar.

Now suppose that both A and B have probability smaller than $1/2$. Let $\tau := \mathbb{M}(d(x, A))$ be the median of $d(x, A)$, and let $C := \{x \in \Lambda : d(x, A) \geq \tau\}$, and $D := \{x \in \Lambda : d(x, A) \leq \tau\}$. Then $\mathbb{P}(C) \geq 1/2$ and $\mathbb{P}(D) \geq 1/2$, moreover it is easy to see that $0 < \tau < d(A, B)$, and $d(A, C) \geq \tau$, and $d(B, D) \geq d(A, B) - \tau$. Therefore using the same argument as in the previous section on A, C and B, D, respectively, we can deduce that

$$\mathbb{P}(A) \leq 2\exp(-\tau^2/C), \text{ and } \mathbb{P}(B) \leq 2\exp(-(d(A, B) - \tau)^2/C),$$

thus

$$\mathbb{P}(A)\mathbb{P}(B) \leq 4\exp(-d(A,B)^2/(2C)).$$

For most of the applications, the functional form is more useful. In this thesis, we will state our inequalities in the functional form. In the next section, we are going to give some examples of the concentration of measure phenomenon.

2.2 Selected examples for concentration

2.2.1 Hoeffding and Bernstein inequalities for sums

The Hoeffding and Bernstein inequalities are the two most frequently used concentration bounds for sums of random variables.

Bernstein's inequality first appeared in Bernstein (1924), and was later rediscovered several times in the literature. Hoeffding's inequality (essentially a special case of Bernstein's inequality, up to constant factors) appeared in Hoeffding (1963), and was generalised to martingales in Azuma (1967).

Let $X_1, \ldots, X_n$ be independent random variables satisfying that $a_i \leq X_i \leq b_i$ for $1 \leq i \leq n$. Then (a simple form of) Hoeffding's inequality states that for any $t \geq 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}(X_i)\right| \geq t\right) \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{2.2.1}$$

Alternatively, assume that $X_1, \ldots, X_n$ are independent random variables satisfying that $|X_i - \mathbb{E}(X_i)| \leq C$ almost surely. Then (a simple form of) Bernstein's inequality

states that for any $t \geq 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}(X_i)\right| \geq t\right) \leq 2\exp\left(\frac{-t^2}{2\sum_{i=1}^{n} \mathrm{Var}(X_i) + (2/3)Ct}\right), \qquad (2.2.2)$$

and the same bound holds for the lower tail.

The advantage of Bernstein's inequality is that it takes into account the variances, whereas Hoeffding's inequality only takes into account the extremal values of the random variables. Therefore it is typically sharper than Hoeffding's inequality.

In the special case when $X_1, \ldots, X_n$ are i.i.d., these inequalities can be thought as a non-asymptotic form of the law of large numbers.

Indeed, in statistics, they can be used in assessing the quality of the estimator $\left(\sum_{i=1}^{n} X_i\right)/n$ of $\mathbb{E}(X_i)$ (see page 65 of Wasserman (2004)). It can be though as a strong form of consistency result for the estimator, in the sense that it not only states that it converges as the sample size tends to infinity, but also gives an explicit error bound for finite sample sizes. We give another application of Hoeffding's inequality in Section 2.2.2.

2.2.2 An application: Quicksort, a randomised algorithm

Quicksort is one of the most efficient sorting algorithms, for sorting a sequence of numbers $x_1, \ldots, x_n$ into increasing order. It is a randomised algorithm, i.e. the time it takes is random, but using concentration inequalities, we are going to show that with high probability, it takes $\mathcal{O}(n\log(n))$ operations. The following exposition is based on Section 2.4 of Dubhashi and Panconesi (2009).

The idea of the algorithm is the following. First, we choose a number out of $x_1, \ldots, x_n$ uniformly, that is, each one with $1/n$ probability. We call this number

the pivot, denoted by p. Then we partition the rest into two blocks, the first block containing the numbers that are less or equal to p, and the second block containing the numbers that are larger than p. This way we obtain a sequence of the form $y_1, \ldots, y_i, p, z_1, \ldots, z_j$, with $y_1, \ldots, y_i$ are smaller or equal to p, and $z_1, \ldots, z_j$ are larger than p (one of these two sets may be empty). Finally, we repeat the same step on $y_1, \ldots, y_i$, and $z_1, \ldots, z_j$ (i.e. the algorithm is recursive).

Now to evaluate how many operations does this algorithm takes, we can notice that the natural way to describe it is by a binary tree. In the root, we put $x_1, \ldots, x_n$, then in each step, the two children of the node become the two sequences $y_1, \ldots, y_i$, and $z_1, \ldots, z_j$. Then there will be a single number on the leaves.

Now since partitioning takes linear time, and every level of the tree contains at most n numbers in total, it is enough to estimate the height of the tree to bound the running time of the algorithm.

Denote the height of the tree by H, then the following proposition gives a bound on it.

Proposition 2.2.1. *For the above algorithm, we have*

$$\mathbb{P}(H \geq 21 \log_2(n)) \leq \frac{1}{n}.$$

Proof. Denote the length of the path from the root to each of the leaves by $P_1, \ldots, P_l$, with $l \leq n$ denoting the total number of leaves. Then $H = \max_{1 \leq i \leq l} P_i$.

Now for $1 \leq i \leq l$, P_i is a random variable, which depends on the choices of pivots in each step of the algorithm. We say that a pivot is good if both of the partitions are at least $1/3$ of the size of the original length, and bad otherwise. Then the length of the sequence after each pivot decreases to less than its two thirds, so the number

of good pivots along any path cannot exceed $\log_{3/2}(n) \leq 2\log_2(n)$. Suppose that a path to a leaf from the root is at least $21\log_2(n)$ long, then among the first $21\log_2(n)$ choices, we must have chosen at most $2\log_2(n)$ good pivots.

Now the probability of choosing a good pivot is $1/3$. If we denote by $Z_1, \ldots, Z_{21\log_2(n)}$ i.i.d. Bernoulli random variables with parameter $1/3$, then $\sum_{i=1}^{21\log_2(n)} \mathbb{E}(Z_i) = 7\log_2(n)$, and thus using (2.2.1) (Hoeffding's inequality), we obtain

$$\mathbb{P}(P_i \geq 21\log_2(n)) \leq \mathbb{P}\left(\sum_{i=1}^{21\log_2(n)} Z_i \leq \sum_{i=1}^{21\log_2(n)} \mathbb{E}(Z_i) - 5\log_2(n)\right)$$
$$\leq \exp(-2(5^2\log_2(n)^2)/(21\log_2(n))) \leq 1/n^2.$$

Now using the union bound, we obtain the result of the proposition. □

A sharper bound on the running time of this algorithm can be obtained using martingale methods, see Section 7.6 of Dubhashi and Panconesi (2009).

There are many other examples in the computer science literature of application of concentration inequalities to estimate the running times of randomised algorithms. For an accessible treatment, we recommend Dubhashi and Panconesi (2009), and Mitzenmacher and Upfal (2005). A related approach is the so called probabilistic method of Erdős, which consists of introducing probability into problems of discrete mathematics that have nothing to do with probability in their original form. Amongst other things, concentration bounds can be used to obtain existence results. For a wonderful exposition of this topic, see Alon and Spencer (2008). Recently, this line of argument has been applied to prove existence results in quantum information theory (see Ahlswede and Winter (2002a), Ahlswede and Winter (2003)).

2.2.3 The bounded differences inequality

The bounded differences inequality is actually a consequence of the Azuma-Hoeffding inequality (due to Azuma (1967), see Section 2.3.1). It became popular after the publication McDiarmid (1989), which has given several interesting applications to this inequality. Since then, the literature calls this result McDiarmid's bounded differences inequality.

Let $X = (X_1, \ldots, X_n)$ be a vector of independent random variables taking values in $\Omega := \Omega_1 \times \ldots \times \Omega_n$, and $f : \Omega \to \mathbb{R}$ be a function satisfying that for some positive constants $c_1, \ldots, c_n$,

$$|f(x) - f(y)| \leq \sum_{i=1}^{n} c_i \cdot \mathbb{1}[x_i \neq y_i] \text{ for every } x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n).$$

Then the bounded differences inequality states that for any $t \geq 0$,

$$\mathbb{P}(f(X) - \mathbb{E}(f) \geq t), \ \mathbb{P}(f(X) - \mathbb{E}(f) \leq -t) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right). \tag{2.2.3}$$

One of the classical applications of this result is proving concentration for the chromatic number of an Erdős-Rényi graph. Let $G(n, p)$ be an Erdős-Rényi graph with edges $X = (X_{i,j})_{1 \leq i < j \leq n}$ being i.i.d. Bernoulli random variables with parameter p. The chromatic number number of the graph, denoted by $\chi(X)$, is the minimal number of colors needed to color the vertices of the graph such that no two vertices of the same color are connected by any edge.

We define $Y_1 := (X_{1,2}, \ldots, X_{1,n}), Y_2 := (X_{2,3}, \ldots, X_{2,n}), \ldots, Y_{n-1} := (X_{n-1,n})$. Then $Y = (Y_1, \ldots, Y_{n-1})$ is just a repartition of X, thus we can define a function χ' such that $\chi'(Y) = \chi(X)$ almost surely. Now it is easy to verify that changing

the value of Y_i can change the chromatic number at most by 1. This means that χ' satisfies the conditions of the bounded differences inequality with $c_1 = \ldots = c_{n-1} = 1$, and thus for any $t \geq 0$,

$$\mathbb{P}(\chi(X) - \mathbb{E}(\chi) \geq t), \ \mathbb{P}(\chi(X) - \mathbb{E}(\chi) \leq -t) \leq \exp\left(\frac{-2t^2}{n-1}\right).$$

The beauty of this result lies in the fact that the chromatic number is a very complicated function of X, and there are no results in the literature about the asymptotic distribution of $\chi(X) - \mathbb{E}(\chi)$. Despite this, the bounded differences inequality gives an elegant way to bound the tails of $\chi(X) - \mathbb{E}(\chi)$.

2.2.4 Talagrand's convex distance inequality

Talagrand's convex distance inequality is a fundamental result that allows to obtain better bounds than those possible using the bounded differences inequality in many examples. The inequality was first stated in the original paper Talagrand (1995). There are several ways to state this inequality. The original proof is based on the set distance formalism (explained in more detail in Section 2.3.2), which then implies concentration for functions. Here we state the form that is most useful for applications, called the *method of non-uniformly bounded differences*. This form of the inequality was first stated in Steele (1997), which also includes several interesting applications.

Theorem 2.2.2. *Let $X = (X_1, \ldots, X_n)$ be a vector of independent random variables taking values in $\Omega := \Omega_1 \times \ldots \times \Omega_n$, and $f : \Omega \to \mathbb{R}$ be a function satisfying that for*

some positive functions $c_1, \ldots, c_n : \Omega \to \mathbb{R}$,

$$f(x) - f(y) \leq \sum_{i=1}^{n} c_i(x) \cdot \mathbb{1}[x_i \neq y_i] \text{ for every } x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n),$$

and $\sum_{i=1}^{n} c_i^2(x) \leq C$ *uniformly in* $x \in \Omega$, *then for any* $t \geq 0$,

$$\mathbb{P}(f(X) - \mathbb{M}(f) \geq t), \mathbb{P}(f(X) - \mathbb{M}(f) \leq -t) \leq \exp\left(\frac{-t^2}{4C}\right),$$

where $\mathbb{M}(f)$ *denotes the median of* $f(X)$.

Remark 2.2.3. This is the classical form of this theorem. In the manuscript Paulin (2014), based on the transportation cost inequality approach of Samson (2000), we improve this result and shown that under the same conditions,

$$\mathbb{P}(f(X) - \mathbb{E}(f) \geq t), \mathbb{P}(f(X) - \mathbb{E}(f) \leq -t) \leq \exp\left(\frac{-t^2}{2C}\right).$$

We do not include the proof in this thesis because of space considerations.

One of the important applications of this result is to show concentration for eigenvalues of random matrices with bounded entries.

Proposition 2.2.4. *Suppose that* $X = (X_{i,j})_{1 \leq i,j \leq n}$ *is a real valued symmetric matrix with* $X_{i,j} = X_{j,i}$ *for every* $1 \leq i, j \leq n$, *and* $(X_{i,j})_{1 \leq i < j \leq n}$ *are independent random variables that satisfy that* $|X_{i,j}| \leq 1$. *Denote the eigenvalues of the matrix* X *by* $\lambda_1(X) \geq \ldots \geq \lambda_n(X)$. *Then for any* $1 \leq s \leq n$, *for any* $t \geq 0$,

$$\mathbb{P}(|\lambda_s(X) - \mathbb{M}(\lambda_s)| \geq t) \leq 4\exp(-t^2/(32s^2)),$$

and the same bound holds for $\lambda_{n-s+1}(X)$.

This proposition is due to Alon, Krivelevich, and Vu (2002). A sharper version (s^2 replaced by s) was obtained in Meckes (2004), also using Talagrand's convex distance inequality.

2.2.5 Gromov-Lévy inequality for concentration on a sphere

Let $S^n \subset \mathbb{R}^{n+1}$ be the surface of an n+1 dimensional sphere of radius 1. Let $f : S^n \to \mathbb{R}$ be a function that is 1-Lipschitz with respect to the geodesic distance on S^n. Let μ be the uniform distribution on S^n, and $X \sim \mu$, then for any $t \geq 0$,

$$|\mathbb{P}(f(X) - \mathbb{M}(f)| \geq t) \leq 4\exp(-(n-1)t^2/2),$$

where $\mathbb{M}(f)$ denotes the median of f. In this form, the result is due to Lévy. It was extended to manifolds with strictly positive Ricci-curvature by Gromov. Recently, this result has found impressive applications in quantum information theory, shedding light on basic properties of entanglement, see Hayden, Leung, and Winter (2006).

2.3 Methods to prove concentration

In this section, we are going to review some of the most popular methods in the literature for proving concentration inequalities. At the time of the writing of this thesis, the field of concentration inequalities has grown very large, with contributions from various areas of mathematics (functional analysis, geometry, probability, statistics, computer science). There is an infinite variety of dependence structures that can arise between random variables. Therefore we make no claim of completeness here,

there are other approaches in the literature, and for some specific types of dependence, they may yield sharper results than those discussed here. However, we have made an effort to explain the basics of those methods that we know to be related to this thesis, and describe their relation to our new results here.

2.3.1 Martingale-type approaches

Martingale-type approaches have been popular for proving concentration inequalities since the classical result of Azuma and Hoeffding (Azuma (1967), Hoeffding (1963)).

Theorem 2.3.1 (Azuma-Hoeffding inequality)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose that $\emptyset = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_{F_n} = \mathcal{F}$ is a filtration of σ-fields. Suppose that $X_0, X_1, \ldots, X_n$ is a martingale with respect to this filtration. Let $D_i := \operatorname{ess\,sup}|X_i - X_{i-1}|$, and let $D^2 := \sum_{i=1}^n D_i^2$, then for $X \sim \mathbb{P}$, for any $t \geq 0$,*

$$\mathbb{P}(|X_n - \mathbb{E}(X_n)| \geq t) \leq 2\exp(-t^2/(2D^2)). \tag{2.3.1}$$

Remark 2.3.2. The upper tail also holds for super-martingales (and symmetrically, the lower tail holds for sub-martingales).

Proof. The proof is based on bounding the moment generating of X_n, and then using Markov's inequality (this argument is standard, and will be used for every concentration bound in this thesis). For $\theta \in \mathbb{R}$, we have

$$\mathbb{E}\left(e^{\theta X_n}\right) = \mathbb{E}\left(e^{\theta \sum_{i=1}^n (X_i - X_{i-1})}\right)$$

Now using Yensen's inequality for convex functions, for any $\theta \in \mathbb{R}$, and $-1 \leq u \leq 1$,

$$e^{\theta u} \leq \frac{1+u}{2}e^{\theta} + \frac{1-u}{2}e^{-\theta}.$$

Now $\mathbb{E}(X_n - X_{n-1}|\mathcal{F}_{n-1}) = 0$, so this means that

$$\begin{aligned}\mathbb{E}\left(e^{\theta(X_n - X_{n-1})}|\mathcal{F}_{n-1}\right) &\leq \mathbb{E}\left(e^{\theta D_n}\cdot\frac{1+\frac{X_n - X_{n-1}}{D_n}}{2} + e^{-\theta D_n}\cdot\frac{1-\frac{X_n - X_{n-1}}{D_n}}{2}\middle|\mathcal{F}_{n-1}\right)\\ &\leq \cosh(\theta D_n) \leq \exp(\theta^2 D_n^2/2).\end{aligned}$$

Now returning to the moment generating function, we can successfully condition on $\mathcal{F}_{n-1}, \mathcal{F}_{n-2}, \ldots, \mathcal{F}_0$, to get

$$\begin{aligned}\mathbb{E}\left(e^{\theta\sum_{i=1}^{n}(X_i - X_{i-1})}\right) &= \mathbb{E}\left(e^{\theta\sum_{i=1}^{n-1}(X_i - X_{i-1})}\cdot\mathbb{E}\left(e^{\theta(X_n - X_{n-1})}|\mathcal{F}_{n-1}\right)\right)\\ &\leq \mathbb{E}\left(e^{\theta\sum_{i=1}^{n-1}(X_i - X_{i-1})}\right)\cdot\exp(\theta^2 D_n^2/2)\\ &\leq \mathbb{E}\left(e^{\theta\sum_{i=1}^{n-2}(X_i - X_{i-1})}\right)\cdot\exp(\theta^2(D_{n-1}^2 + D_n^2)/2) \leq \ldots \leq \exp(\theta\mathbb{E}(X_n))\cdot\exp(\theta^2 D^2/2).\end{aligned}$$

Now we can use Markov's inequality to obtain the concentration bounds. For any $t \geq 0$, $\theta \geq 0$, $\mathbb{E}(e^{\theta X_n - \mathbb{E}(X_n)}) \geq \mathbb{P}(X_n - \mathbb{E}(X_n) \geq t)\cdot\exp(\theta t)$, and $\mathbb{E}(e^{\theta X_n - \mathbb{E}(X_n)}) \leq \exp(\theta^2 D^2/2)$, thus

$$\mathbb{P}(X_n - \mathbb{E}(X_n) \geq t) \leq \exp(\theta^2 D^2/2 - \theta t).$$

Now optimising in θ shows that the minimum of the right hand side is taken at $\theta = t/D^2$, and thus we obtain $\mathbb{P}(X_n - \mathbb{E}(X_n) \geq t) \leq \exp(-t^2/(2D^2))$. The proof of the lower tail is similar (using negative values of t and θ). □

This theorem implies McDiarmid's bounded differences inequality for independent random variables. In Section 3.2 of Chapter 3, we are going to generalise this proof to Markov chains, and show a version of the bounded differences inequality with constants depending on the mixing time of the chain.

The martingale method was used to prove concentration for Hamming Lipschitz functions of uniform permutations in Maurey (1979) (see also Corollary 4.3 of Ledoux (2001)). It has also been generalised to apply to some non-Lipschitz functions, in particular, multivariate polynomials, in Vu (2002), and Kim and Vu (2000). Such bounds have been applied, for example, to the number of triangles in the Erdős-Rényi graph.

Combining coupling ideas with martingale arguments has proven fruitful for proving concentration inequalities for dependent variables, see Külske (2003), and Chazottes, Collet, Külske, and Redig (2007).

2.3.2 Talagrand's set distance method

Let $\Omega = \Omega_1 \times \ldots \times \Omega_n$. For a vector $\alpha \in \mathbb{R}^n_+$, we define the distance $d_\alpha : \Omega^2 \to \mathbb{R}$ as $d_\alpha(x, y) = \sum_{i=1}^n \alpha_i \mathbb{1}[x_i \neq y_i]$. We define Talagrand's convex distance between a set $A \subset \Omega$ and a point $x \in \Omega$ as

$$d_T(x, A) := \sup_{\alpha \in \mathbb{R}^n_+, \sum \alpha_i^2 \leq 1} \inf_{y \in A} d_\alpha(x, y). \tag{2.3.2}$$

Then the strongest form of Talgrand's convex distance inequality for product spaces is the following.

Theorem 2.3.3. *Let μ be a product measure on Ω. Let $X \sim \mu$. Then for any $A \subset \Omega$,*

$$\mathbb{E}\left(e^{d_T^2(X,A)/4}\right) \leq \frac{1}{\mathbb{P}(A)}. \tag{2.3.3}$$

The original proof of this result is based on mathematical induction in the dimension n.

Let $\overline{A}_t := \{x \in \Omega : d_T(x, A) > t\}$, then this theorem implies the following weaker form. For any $t \geq 0$, any $A \subset \Omega$,

$$\mathbb{P}(A) \cdot \mathbb{P}(\overline{A}_t) \leq \exp(-t^2/4). \tag{2.3.4}$$

This, in turn, implies the method of non-uniformly bounded differences (Theorem 2.2.2). For a short proof of these, see pages 139-140 of Dubhashi and Panconesi (2009).

Besides product spaces, Talagrand's convex distance inequality also holds for uniform permutations (see Talagrand (1995)). In this case, an equation of the form of (2.3.3) holds, with constant 16 instead of 4.

In addition to the definition (2.3.2), Talagrand has defined set distances in several other ways as well. His so called "control by several points method" generalises $d_T(x, A)$ to define a distance between a point and several sets, of the type $d_T(x, A_1, \ldots, A_q)$. This method has lead to important new concentration inequalities for suprema of empirical processes in product spaces (in particular, for sums of the form $\sup_{f \in \mathcal{F}} \sum_{i=1}^n f(X_i)$, where $X_1, \ldots, X_n$ are independent random variables, and $\mathcal{F}$ is a countable set of real valued functions). These inequalities have proven to be very useful for applications in model selection, and machine learning.

For a concise proof of this result Talagrand's inequality for uniform permutations, see Section 8.2 of Ledoux (2001). Talagrand's inequality for uniform permutations was further generalised in McDiarmid (2002), and Luczak and McDiarmid (2003). Boucheron, Bousquet, and Lugosi (2005a) is a great survey on applications of concentration inequalities for empirical processes to the theory of classification. For applications to model selection problems, see Massart (2007).

Finally, it is worth noting that most of the inequalities obtained by Talagrand's set distance method have been also proven using Ledoux's log-Sobolev-type entropy method (Ledoux (1995/97), Massart (2000)), and using transportation cost inequalities (see Dembo (1997)).

2.3.3 Log-Sobolev inequalities and the entropy method

In this section first we state the simplest form of log-Sobolev inequalities, show how they imply concentration via the so called Herbst argument. Then we explain the basics of the entropy method.

Log-Sobolev inequalities were introduced in Gross (1975) in relation with quantum field theory. They have later found applications in many fields of mathematics, see the lecture notes Guionnet and Zegarlinksi (2003), and Ané, Blachère, Chafaï, Fougères, Gentil, Malrieu, Roberto, and Scheffer (2000). For applications to Markov chains (bounding for the spectral gap of the chain), see Diaconis and Saloff-Coste (1996). More recently, a version of log-Sobolev inequalities, the entropy method, has proven to be a powerful method to prove concentration inequalities (see Boucheron, Lugosi, and Massart (2013b)).

Given a probability space $(\Omega, \mathcal{F}, \mu)$, and a measurable function $f : \Omega \to \mathbb{R}$, we

define its entropy as

$$\mathrm{Ent}_\mu(f) := \mathbb{E}_\mu(f\log(f)) - \mathbb{E}_\mu(f)\log(\mathbb{E}_\mu(f)).$$

Now in the case of $\Omega = \mathbb{R}^n$, and $\mathcal{F}$ being all the Borel sets of $\mathbb{R}^n$, we say that μ satisfies the log-Sobolev inequality with constant C if for all smooth enough functions f,

$$\mathrm{Ent}_\mu\left(f^2\right) \leq 2C\mathbb{E}_\mu\left(|\nabla f|^2\right), \tag{2.3.5}$$

with $|\nabla f(x)|$ denoting the Euclidean length of the gradient vector of f at point x.

Then the following theorem gives an example about log-concave distributions where the log-Sobolev constant C can be bounded.

Theorem (Theorem 5.2 of Ledoux (2001))**.** *Suppose that $\Omega = \mathbb{R}^n$, and $\mathcal{F}$ contains all the Borel sets. Let $\mathrm{d}\mu = e^{-U(x)}\mathrm{d}x$, where for some $c > 0$, $\lambda_{\min}\left(\mathrm{Hess}\, U(x)\right) \geq c$ uniformly for every $x \in \mathbb{R}^n$ ($\lambda_{\min}$ denotes the smallest eigenvalue). Then for all smooth enough functions f on $\mathbb{R}^n$,*

$$\mathrm{Ent}_\mu\left(f^2\right) \leq \frac{2}{c}\mathbb{E}_\mu\left(|\nabla f|^2\right),$$

i.e. the log-Sobolev inequality holds with constant $C = 1/c$.

Remark 2.3.4. In the special case of the n dimensional standard Gaussian distribution, $U(x) = \|x\|_2^2/2 + n/2\log(2\pi)$, and thus $\lambda_{\min}\left(\mathrm{Hess}\, U(x)\right) = \lambda_{\min}\left(\boldsymbol{I}\right) = 1$, therefore we have $c = 1$ and $C = 1$.

The following proposition relates the log-Sobolev inequality with concentration of Lipschitz functions.

Proposition 2.3.5 (Herbst argument)**.** *Suppose that μ satisfies* (2.3.5)*. Let $X \sim \mu$, the for any $f : \mathbb{R}^n \to \mathbb{R}$, any $t \geq 0$,*

$$\mathbb{P}(|f(X) - \mathbb{E}(f)| \geq t) \leq 2 \exp\left(-\frac{t^2}{2C\|f\|_{\text{Lip}}^2}\right),$$

where $\|f\|_{\text{Lip}}$ denotes the Lipschitz coefficient of f with respect to the Euclidean distance.

Remark 2.3.6. The proof of this result is given on pages 94-95 of Ledoux (2001).

In the special case of the standard normal distribution, $C = 1$, thus we obtain the Cirelson-Ibragimov-Sudakov inequality (see Section 1.2.1 of Massart (2007)).

Proposition 2.3.7. *Let $X = (X_1, \ldots, X_n)$ be a vector of independent standard normal random variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a 1-Euclidean Lipschitz function. Then for any $t \geq 0$,*

$$\mathbb{P}(|f(X) - \mathbb{E}(f)| \geq t) \leq 2 \exp(-t^2/2).$$

Now we turn to the basics of the entropy method.

A classical inequality from probability theory is the Efron-Stein inequality (in this form, see Boucheron, Lugosi, and Massart (2003)).

Theorem 2.3.8. *Let $Z = g(X_1, \ldots, X_n)$ be square integrable, where $X_1, \ldots, X_n$ are independent random variables. Let $X_1', \ldots, X_n'$ be independent copies of them. For some real valued function g, let $Z := g(X_1, \ldots, X_n)$, and $Z^{(i)} := g(X_1, \ldots, X_i', \ldots, X_n)$. Then*

$$\text{Var}(Z) \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}(Z - Z^{(i)})^2,$$

whenever all the expectations exist.

The advantage of this result is that it is using the typical deviations of g when changing each of the random variables $X_1, \ldots, X_n$ separately (instead of using the maximal possible deviations, as in the bounded differences inequality). The disadvantage is that it only gives bound on the variance, and not an exponential concentration result. The entropy method allows us to recover exponential concentration bounds of similar type.

The following theorem is an exponential version of the Efron-Stein inequality (see Boucheron, Lugosi, and Massart (2003)).

Theorem 2.3.9. *Let $X_1, \ldots, X_n$, Z, and $Z^{(i)}$ be as in Theorem 2.3.8. Let*

$$V_+ := \mathbb{E}\left[\sum_{i=1}^{n}(Z - Z^{(i)})^2 \mathbb{1}[Z > Z^{(i)}]|X_1, \ldots, X_n\right], \text{ and}$$

$$V_- := \mathbb{E}\left[\sum_{i=1}^{n}(Z - Z^{(i)})^2 \mathbb{1}[Z < Z^{(i)}]|X_1, \ldots, X_n\right].$$

Then for all $\theta > 0$ and $\lambda \in (0, 1/\theta)$,

$$\log \mathbb{E}[\lambda(Z - \mathbb{E}(Z))] \leq \frac{\lambda\theta}{1 - \lambda\theta} \log \mathbb{E}\left[\exp\left(\frac{\lambda V_+}{\theta}\right)\right], \text{ and}$$
$$\log \mathbb{E}[-\lambda(Z - \mathbb{E}(Z))] \leq \frac{\lambda\theta}{1 - \lambda\theta} \log \mathbb{E}\left[\exp\left(\frac{\lambda V_-}{\theta}\right)\right].$$

These inequalities give bounds on the moment generating function of $Z - \mathbb{E}(Z)$ in terms of the moment generating function of V_+ and V_-. The mean of V_+ and V_- can be expressed as

$$\mathbb{E}(V_+) = \mathbb{E}(V_-) = \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - Z^{(i)}\right)^2\right],$$

which is exactly the bound from the Efron-Stein inequality. If we assume that V_+ has finite exponential moments for a non-empty range of positive exponents, then it follows from the theorem that for small values of λ, $\log \mathbb{E}[\lambda(Z - \mathbb{E}(Z))] \leq \exp(\lambda^2 \mathbb{E}(V_+))$, which in turn implies that for sufficiently small deviations, Gaussian tails hold with constants proportional to the right hand side of the Efron-Stein inequality, $\sum_{i=1}^n \mathbb{E}(Z - Z^{(i)})^2$. Thus whenever the Efron-Stein bound gives the right order of variance, we can get sharp Gaussian tails for sufficiently small deviations.

The proof of Theorem 2.3.9 is based on the following modified log-Sobolev inequality (see Massart (2000)).

Theorem 2.3.10. *Let $\psi(x) := e^x - x - 1$. Suppose that $X_1, \ldots, X_n$ are independent random variables, and $X_1', \ldots, X_n'$ are independent copies of them. For some real valued function g, let $Z := g(X_1, \ldots, X_n)$, and $Z_i' := g(X_1, \ldots, X_i', \ldots, X_n)$. Then for any $s > 0$,*

$$s\mathbb{E}[Ze^{sZ}] - \mathbb{E}[e^{sZ}] \log \mathbb{E}[e^{sZ}] \leq \sum_{i=1}^n \mathbb{E}[e^{sZ} \psi(-s(Z - Z_i'))].$$

Moreover, denote $\tau(x) := x(e^x - 1)$. Then for all $s \in \mathbb{R}$,

$$s\mathbb{E}[Ze^{sZ}] - \mathbb{E}[e^{sZ}] \log \mathbb{E}[e^{sZ}] \leq \sum_{i=1}^n \mathbb{E}\left[e^{sZ} \tau(-s(Z - Z_i'))\mathbb{1}[Z > Z_i']\right], \text{ and}$$
$$s\mathbb{E}[Ze^{sZ}] - \mathbb{E}[e^{sZ}] \log \mathbb{E}[e^{sZ}] \leq \sum_{i=1}^n \mathbb{E}\left[e^{sZ} \tau(-s(Z_i' - Z))\mathbb{1}[Z < Z_i']\right].$$

The entropy method was shown to imply the strongest form (2.3.3) of Talagrand's convex distance inequality in Boucheron, Lugosi, and Massart (2009). In Chapter 7, we use parts of the approach of that paper to prove Talagrand's convex distance

inequality for dependent variables. The entropy method has also been generalised to obtain moment bounds for functions of independent random variables in Boucheron, Bousquet, Lugosi, and Massart (2005b).

2.3.4 Transportation cost inequality method

Transportation cost inequalities are a powerful tools of proving concentration results. They were introduced by Marton, based on ideas from information theory. Here we briefly review the basics of this method.

Suppose that we have a Polish metric space (Ω, d), and distributions μ and ν on it. Then the L^1 and L^2 *Wasserstein distances* are defined as

$$W_1(\mu,\nu) := \inf_{\pi[X\sim\mu, Y\sim\nu]} \mathbb{E}_\pi(d(X,Y)), \tag{2.3.6}$$

$$W_2(\mu,\nu) := \inf_{\pi[X\sim\mu, Y\sim\nu]} [\mathbb{E}_\pi(d^2(\mu,\nu))]^{1/2}, \tag{2.3.7}$$

where the infimum is taken over all distributions π defined on Ω^2 having marginals μ and ν. Define the *relative entropy* of two measures ν and μ as

$$D(\nu||\mu) = \int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right) \mathrm{d}\nu, \tag{2.3.8}$$

with the convention that it is infinity if ν is not absolutely continuous with respect to μ. A distribution μ on (Ω, d) is said to satisfy a *transportation cost inequality* with constant c if for any distribution ν on (Ω, d),

$$W_1(\nu,\mu) \leq \sqrt{2cD(\nu||\mu)}.$$

Alternatively, a distribution μ on (Ω, d) is said to satisfy a *quadratic transportation cost inequality* with constant c if for any distribution ν on (Ω, d),

$$W_2(\nu, \mu) \leq \sqrt{2cD(\nu||\mu)}.$$

In general spaces, transportation cost inequalities imply Gaussian concentration for d-Lipschitz functions (in fact, as it was shown in Djellout, Guillin, and Wu (2004), Gaussian concentration is equivalent to transportation cost inequalities). In product-like spaces (such that independent random variables, or uniform permutations) they can be shown to imply McDiarmid's bounded differences inequality.

Quadratic transportation cost inequalities are stronger results. In product-like spaces, some special type of quadratic transportation cost inequalities also imply Talagrand's convex distance inequality, Bernstein's inequality, and further inequalities, see Samson (2000), Marton (2003). In the seminal work Otto and Villani (2000), it was shown that in a general setting, log-Sobolev inequalities imply quadratic transportation cost inequalities.

One great success of the transportation cost inequality method was proving concentration inequalities for so called contracting Markov chains. For a homogeneous Markov chain with Polish state space Ω, and transition probabilities $P(x, y)$, let us denote $a := \sup_{x,y\in\Omega} d_{\mathrm{TV}}(P(x, \cdot), P(y, \cdot))$, then Proposition 1 of Marton (1996b) proves a transportation cost inequality $1/(1 - a)^2$ times worse than in the independent case (see (3.1.1) for the definition of the total variational distance). Marton (1996a) shows a quadratic transportation cost inequality for such chains, again, with constants $1/(1 - a)^2$ times weaker than in the independent case. Further extension was given in Samson (2000) and an unpublished manuscript of Marton.

In this thesis, we improve upon these bounds for Markov chains, and show that McDiarmid's bounded difference inequality holds with constants that are the mixing time of the chain times weaker than in the independent case. In fact, we have found two proofs for this result, one using transportation cost inequalities (which is more general, and also yields Talagrand's convex distance inequality, Bernstein's inequality, and further inequalities), and one simpler approach using a martingale-type argument. Because of space considerations, we have decided to only include the martingale-type approach in this thesis.

In this short paragraph, we have only attempted a cover the basics of the transportation cost inequality method, which have became popular in the last decade, and found many connections with other fields. More complete references are Villani (2009), and Gozlan and Léonard (2010).

2.3.5 Spectral methods

For sums of the form $f(X_1)+\ldots+f(X_n)$, where $X_1, \ldots, X_n$ is a Markov chain, spectral methods can be used to obtain variance and concentration bounds. For reversible chains, these methods take into account the spectrum of Markov kernel, in particular, they depend on its spectral gap (the distance between its largest eigenvalue, 1, and its second largest eigenvalue).

The first Hoeffding-type concentration bound, in the case when f is a 0-1 valued indicator function, was given in Gillman (1998) (see also Kahale (1997) for a sharper version). This bound have used the perturbation theory of linear operators. In fact, much earlier, asymptotic bounds have been obtained for such sums using this theory in Nagaev (1957).

Building upon the ideas of Gillman (1998), and also using Kato's perturbation theory of linear operators, Lezaud (1998b) has proven Bernstein-type concentration bounds for reversible, and non-reversible Markov chains, and processes. In the reversible case, the bound depends on the spectral gap of the chain, while in the non-reversible case, the spectral gap of its multiplicative reversiblication ($\boldsymbol{P}^*\boldsymbol{P}$).

A sharp version of Hoeffding's inequality was proven in León and Perron (2004) for reversible Markov chains using a stochastic ordering-type argument.

In Section 3.3 of Chapter 3, we improve upon these results, using the same approach as Lezaud (1998b), but with more careful estimation. For reversible chains, we give Bernstein bounds as a function of the asymptotic variance, while for non-reversible chains, as a function of the pseudo spectral gap of the chain (a generalisation of the multiplicative reversiblication).

2.3.6 Semigroup tools, and the coarse Ricci curvature

Semigroup arguments can be used to obtain concentration inequalities for probability measures arising as the stationary distribution of Markov processes. They have been successfully applied to show concentration for spheres, manifolds with strictly positive Ricci curvature (Gromov-Lévy theorem), as well as log-concave densities (such as the Gaussian measure). The main idea of these methods is that we choose a Markov process with analytically simply described generator (such as the heat semigroup, with generator $\boldsymbol{L} = \nabla$, the Laplace operator), and then use various integration by parts formulas to get bounds on the moment generating function $\mathbb{E}(e^{\lambda f})$ for smooth Lipschitz functions f (which then translate into bounds for all Lipschitz functions by limiting arguments). The advantage of these methods is that they can lead to

sharp bounds, and the arguments can be very concise. The disadvantage is that for different types of Markov processes, different tricks need to be used.

A generalisation of the semigroup approach to discrete time Markov chains is the following general concentration inequality (Theorem 3.3 of Ledoux (2001)).

Theorem 2.3.11. *Let $P(x, y)$ be a reversible Markov kernel, with finite state space Ω, stationary distribution π, and spectral gap γ. For a function $f : \Omega \to \mathbb{R}$, define*

$$|||f|||_\infty := \frac{1}{2} \sup_{x \in \Omega} \sum_{y \in \Omega} |f(x) - f(y)|^2 \cdot P(x, y).$$

Let $X \sim \pi$, then for any $t \geq 0$,

$$\mathbb{P}(|f(X) - \mathbb{E}(f)| \geq t) \leq 6 \exp(-t\sqrt{\gamma}/(2|||f|||_\infty),$$

with γ denoting the spectral gap of the chain.

This result is quite general, since it proves concentration for possibly non-Lipschitz functions. However, it only shows exponential bounds. In fact, Gaussian bounds can hold in many cases. Thus it is rarely possible to obtain sharp bounds using this theorem.

Another, more recent approach is the so called coarse Ricci curvature method initiated by Ollivier, which allows to prove concentration inequalities for distributions arising as the stationary distribution of Markov chains. The bounds depends mainly on 4 quantities (the latter three is defined below), the Lipschitz constant of the function, the coarse Ricci curvature, the local dimension, and the diffusion constant. Let $P(x, \mathrm{z})$ be a Markov kernel with Polish metric state space (Ω, d). For any $x, y \in \Omega$,

$x \neq y$, the *coarse Ricci curvature* is defined as

$$\kappa(x,y) = 1 - \frac{W_1(P_x, P_y)}{d(x,y)} \text{ for } x \neq y, \text{ and } \kappa = \sup_{x,y\in\Omega, x\neq y} \kappa(x,y),$$

where P_x denotes the measure $P(x, \mathrm{d}z)$, and W_1 denotes the Wasserstein distance of P_x and P_y (as defined in (2.3.6)). The *local dimension* $n(x)$ is defined as

$$n(x) := \frac{\sigma(x)^2}{\sup\{\mathrm{Var}_{P_x} f, f : \mathrm{Supp}\, P_x \to \mathbb{R} \text{ 1 - Lipschitz}\}}.$$

Then $n(x) \geq 1$, and when Ω is an N dimensional space or the surface of an N dimensional manifold, $n(x)$ is usually related to N. Let $\sigma^2(x)$ is defined as

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

Based on these quantities, Ollivier (2009) shows that for $X \sim \pi$, for any $f : \Omega \to \mathbb{R}$ with Lipschitz coefficient $\|f\|_{\mathrm{Lip}}$, there is some $t_{\max} > 0$ such that for $0 \leq t \leq t_{\max}$,

$$\mathbb{P}(|f(X) - \mathbb{E}_\pi(f)| \geq t) \leq \exp\left(-\frac{t^2}{6\kappa \|f\|_{\mathrm{Lip}}^2 \mathbb{E}_\pi\left(\frac{\sigma^2}{n}\right)}\right). \tag{2.3.9}$$

The value of $t_{\max}$ depends on the maximal diameter of the support of the measure $P_x(\mathrm{d}z)$, and on the Lipschitz coefficient of $\sigma^2(x)/n(x)$.

This method has been successfully applied to numerous examples. In particular, by showing that κ can be lower bounded on manifolds positive Ricci curvature, it recovers the celebrated Gromov-Lévy theorem (up to constant factors). In Chapter 4, we generalise this method by considering the coarse Ricci curvature of several steps in the Markov chain.

2.3.7 Concentration by Stein's method of exchangeable pairs

Stein's method of exchangeable pairs was adapted for proving concentration inequalities by Chatterjee (2005). Here we explain the basics of this method. Suppose that Ω is a Polish space, and $F : \Omega^2 \to \mathbb{R}$ is an antisymmetric function. Suppose that (X, X') is an exchangeable pair taking values in Ω. Let $f(X) := \mathbb{E}(F(X, X')|X)$, and

$$\Delta(X) := \frac{1}{2}\mathbb{E}(|(f(X) - f(X'))F(X, X')||X).$$

The concentration of f around its mean is determined by $\Delta(X)$. We have $\mathrm{Var}(f) \leq \mathbb{E}(\Delta(X))$, and if $\Delta(X) \leq C$ almost surely, then

$$\mathbb{P}(|f(X) - \mathbb{E}(f)| \geq t) \leq 2\exp(-t^2/(2C)).$$

More generally, if $\Delta(X) \leq \varphi(f(X))$ for some function $\varphi(x) \sim x^\alpha$ with $0 \leq \alpha < 2$, then $\mathbb{P}(|f(X) - \mathbb{E}(f)| \geq t) \leq 2\exp(-\mathcal{O}(t^{2-\alpha}))$ holds.

There are several examples of models where a smart choice of $F(X, X')$ can lead to a concentration inequality for an interesting function f (see the references below). On the other hand, the converse problem, how can we find $F(X, X')$ for a given function f, is also important. This problem is addressed in Chapter 4 of Chatterjee (2005). Denote by P the Markov kernel generated by the exchangeable pair (X, X') (that is, $P(g)(x) = \mathbb{E}(g(X')|X = x)$). Then under some technical assumptions ensuring the convergence, F defined as

$$F(x, y) := \sum_{i=0}^{\infty} \left[P^k(f)(x) - P^k(f)(y)\right]$$

is antisymmetric, and satisfies $\mathbb{E}(F(X, X')|X) = f(X)$. This construction is used in Chatterjee (2005) to prove a version of McDiarmid's bounded differences inequality for weakly dependent random variables satisfying the Dobrushin condition. Applications of the method in Chatterjee (2005) include proving mean-field equations for the Curie-Weiss and Sherrington-Kirkpatrick models, pseudo maximal likelihood estimation for the Ising model (see also Chatterjee (2007)). The $\Delta(X) \leq \varphi(f(X))$ case is explored in Chatterjee and Dey (2010), with further applications to statistical physical models, and random graphs.

In Chapter 5, we generalise this method to show Talagrand's convex distance inequality under the Dobrushin condition. In Chapter 4, we explore a connection between this method and Ollivier's coarse Ricci curvature. This allows us to prove concentration for Lipschitz functions beyond the Dobrushin condition case.

Finally, we note that recently other variants of Stein's method, size-biasing and zero-biasing, has also been used to prove concentration inequalities, see Ghosh and Goldstein (2011), Goldstein and Islak (2013).

2.3.8 Janson's trick for sums of dependent random variables

We say that $X_1, \ldots, X_n$ are locally dependent random variables (more specifically, (LD) dependent) if for every X_i there is a collection of random variables A_i such that it is independent from all the rest. Denote by $G = (V, E)$ the dependence graph of these variables, i.e. a graph having vertices $1, \ldots, n$, and edges between vertex i and j if and only if A_i contains X_j or A_j contains X_i. Denote by $\chi(G)$ the chromatic number of the graph, then we can divide $X_1, \ldots, X_n$ into $\chi(G)$ groups $\overline{X}_1, \ldots, \overline{X}_{\chi(G)}$ such that each of the groups contains independent random variables.

Denote by $Y_1, \ldots, Y_{\chi(G)}$ the sum of the random variables in each group, then using Jensen's inequality, we can bound the movement generating function as

$$\mathbb{E}\left(e^{\theta \sum_{i=1}^{n} X_i}\right) = \mathbb{E}\left(e^{\theta \sum_{i=1}^{\chi(G)} Y_i}\right) \leq \frac{1}{k} \sum_{i=1}^{\chi(G)} \mathbb{E}(e^{\theta k Y_i}), \tag{2.3.10}$$

which can be bounded above using independence. More generally, we have

$$\mathbb{E}\left(e^{\theta \sum_{i=1}^{\chi(G)} Y_i}\right) \leq \sum_{i=1}^{\chi(G)} c_i \mathbb{E}(e^{\theta Y_i / c_i})$$

for any positive reals $c_1, \ldots, c_{\chi(G)}$ satisfying $\sum_{i=1}^{\chi(G)} c_i = 1$.

This method has been applied in Janson (2004) to obtain Hoeffding and Bernstein inequalities for sums of locally dependent random variables.

In this thesis, we generalise this trick somewhat further, by noticing that the groups $\overline{X}_1, \overline{X}_1, \ldots$ do not need to consist of independent random variables, it is sufficient if the dependence between the variables in each group is small.

Our definition of the pseudo spectral gap in Chapter 3, Section 3.3 is motivated by this method.

2.3.9 Matrix concentration inequalities

Suppose that $X = (X_1, \ldots, X_n)$ is a vector of random variables, and $f(X_1, \ldots, X_n)$ is a Hermitian matrix valued function. Then in many cases, we are interested in the concentration properties of $f(X)$ around its mean, in the sense that we want to get bounds on the quantity $\mathbb{P}(\|f(X) - \mathbb{E}(f)\| \geq t)$, with $\|\cdot\|$ denoting the L_2 operator norm. Bounds of this type are called *matrix concentration inequalities*. They have

first appeared in quantum information theory (Ahlswede and Winter (2002b)), and became popular after Tropp (2012), which has considerably sharpened the previous results, and proven Azuma-Hoeffding and Bernstein-type inequalities.

The main tool for proving such inequalities is the trace moment generating function, defined as $\mathbb{E}\mathrm{tr}\exp(\theta f(X))$. This function is behaving quite similarly as the moment generating function in the scalar case, and by bounding it, we can obtain a concentration bound for $\mathbb{P}(\|f(X) - \mathbb{E}(f)\| \geq t)$. However, a considerable difficulty in the matrix case is that even for sums of independent random matrices, the trace moment generating function does not factorizes to the product of individual terms (because of the non-commutativity of the matrix product). This difficulty can be solved with the help of various trace inequalities.

In addition to the independent case treated in Tropp (2012), concentration inequalities have been also proven for functions of dependent random variables, using the Stein's method approach of Section 2.3.7. Mackey, Jordan, Chen, Farrell, and Tropp (2012) has introduced the concept of Stein pairs, and used it to show concentration for sums of dependent random matrices. This was further generalised in Paulin, Mackey, and Tropp (2013) to consider more general functions, and a matrix version of McDiarmid's bounded differences inequality was proven for weakly dependent random variables (we do not include it in this thesis because of space considerations).

These inequalities have found numerous applications in statistics (Rohde and Tsybakov (2011)), and computer science, in particular in the field of compressed sensing (Tropp (2011), Candès and Davenport (2013)).

2.3.10 Other methods

In this section we mention a few other methods for proving concentration inequalities.

The *regeneration times* approach is an important method for proving concentration inequalities, and various limit theorems for Markov chains, by essentially deducing them from results for independent random variables. Adamczak (2008) and Adamczak and Bednorz (2012) have used this approach to prove Bernstein inequality, and a version of Talagrand's inequality for empirical processes, for Markov chains (see also Douc, Moulines, Olsson, and van Handel (2011) for a concentration bound for sums of functions of Markov chains). Moreover, this chapter also uses *truncation* to prove inequalities for sums of unbounded functions of dependent random variables (a truncation approach was also used in the earlier result van de Geer (2002)). In the Appendix of this thesis, motivated by the regeneration times approach, we show that for sums of unbounded functions of Markov chains, concentration can be much weaker than in the case of independent summands (i.e. the sums of random variables with gaussian tails will not necessarily be gaussian). Moreover, in Section 3.3 of Chapter 3, we state a proposition based on the truncation method, for generalising our results to unbounded summands.

Negative dependence between random variables X and Y typically corresponds to the condition that for any monotone increasing functions f and g, $E(f(X)g(Y)) \leq \mathbb{E}(f(X))\mathbb{E}(g(Y))$. Under this kind of dependence (and further generalisations), Hoeffding and Bernstein type inequalities hold for sums, similarly as in the independent case (the proof is based on factorizing the moment generating function using the negative dependence condition). See Dubhashi and Ranjan (1998) and Section 3 of Dubhashi and Panconesi (2009) for more details, and examples.

Concentration bounds under further, interesting types of dependence structures are proven in Gavinsky, Lovett, Saks, and Srinivasan (2012), and in Unger (2009).

Chapter 3

Concentration for Markov chains[1]

3.1 Introduction

Consider a vector of random variables

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

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

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

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

[1]This chapter is based on the manuscripts Paulin (2014) and Gyori and Paulin (2014).

variables $X_1, \ldots, X_n$ 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 complicated. Unlike limit theorems, these bounds hold non-asymptotically, that is for any fixed n. Our references on concentration inequalities are Ledoux (2001), and Boucheron, Lugosi, and Massart (2013a).

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 chapter, 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 chapter 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 Lezaud (1998a). 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",

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

and prove similar bounds using it. Moreover, we show that just like $1/\gamma$, $1/\gamma_{\text{ps}}$ 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 Chazottes, Collet, Külske, and Redig (2007) 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 chapter.

In Section 3.1.1, we state basic definitions about general state space Markov chains. This is followed by two sections presenting our results. In Section 3.2, we define Marton couplings, a coupling structure introduced in Marton (2003), 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 3.3, 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 3.4, we generalise the bounds of the previous two sections to continuous time Markov processes. We apply our results to obtain a concentration inequality for the average number of customers in an M/M/1 queue. In Section 3.5, we compare our results with the previous inequalities in the literature, and finally Section 3.6 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 Kontorovich (2007).

3.1.1 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 Roberts and Rosenthal (2004). 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

$$d_{\mathrm{TV}}(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|, \tag{3.1.1}$$

or equivalently

$$d_{\mathrm{TV}}(P, Q) := \inf_{(X,Y)} \mathbb{P}(X \neq Y), \tag{3.1.2}$$

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 Lindvall (1992), see also Lemma 5.5.1 for a concrete construction).

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 Goldstein (1978/79). We will call this type of coupling as Goldstein's maximal coupling (which we will define precisely in Proposition 3.2.6). 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 Roberts and Rosenthal (2004))**.** *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 3.1.1. A Markov chain with stationary distribution π, state space Ω, and transition kernel $P(x, dy)$ is *uniformly ergodic* if

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

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

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

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

Remark 3.1.2. 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 Levin, Peres, and Wilmer (2009).

Definition 3.1.3 (Mixing time for time homogeneous chains)**.** Let X_1, X_2, $X_3, \ldots$ 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

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

The fact that $t_{\mathrm{mix}}(\epsilon)$ is finite for some $\epsilon < 1/2$ (or equivalently, t_{mix} is finite) is equivalent to the *uniform ergodicity* of the chain, see Roberts and Rosenthal (2004), Section 3.3. We will also use the following alternative definition, which also works for time inhomogeneous Markov chains.

Definition 3.1.4 (Mixing time for Markov chains without assuming time homogeneity)**.** Let $X_1,\ldots,X_N$ 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

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

Remark 3.1.5. One can easily see that in the case of time homogeneous Markov chains, by triangle inequality, we have

$$\tau(2\epsilon)\le t_{\mathrm{mix}}(\epsilon)\le\tau(\epsilon). \tag{3.1.3}$$

Similarly to Lemma 4.12 of Levin, Peres, and Wilmer (2009) (see also proposition

3.(e) of Roberts and Rosenthal (2004)), one can show that $\overline{d}(t)$ is subadditive

$$\overline{d}(t+s) \leq \overline{d}(t) + \overline{d}(s), \tag{3.1.4}$$

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

$$\tau(\epsilon^k) \leq k\tau(\epsilon), \text{ and thus } t_{\text{mix}}\left((2\epsilon)^k\right) \leq k t_{\text{mix}}(\epsilon). \tag{3.1.5}$$

3.2 Marton couplings

In this section, we are going to prove concentration inequalities using Marton couplings. First, in Section 3.2.1, we introduce Marton couplings (which were originally defined in Marton (2003)), 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.

3.2.1 Preliminaries

In the following, we will consider dependent random variables $X = (X_1, \ldots, X_N)$ 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 $Y = (Y_1, \ldots, Y_N)$ is another random vector taking values in Λ, with distribution Q. We will refer to distribution of a vector $(X_1, \ldots, X_k)$ as $\mathcal{L}(X_1, \ldots, X_k)$, and

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

will denote the conditional distribution of $X_{k+1}, \ldots, X_N$ under the condition $X_1 = x_1, \ldots, X_k = x_k$. Let $[N] := \{1, \ldots, N\}$. We will denote the operator norm of a square matrix Γ by $\|\Gamma\|$. The following is one of the most important definitions of this chapter. It has appeared in Marton (2003).

Definition 3.2.1 (Marton coupling)**.** Let $\mathcal{X} := (\mathcal{X}_1, \ldots, \mathcal{X}_\mathcal{N})$ be a vector of random variables taking values in $\Lambda = \Lambda_1 \times \ldots \times \Lambda_\mathcal{N}$. We define a *Marton coupling* for $\mathcal{X}$ as a set of couplings

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

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

(i) $\mathcal{X}_1^{(x_1,\ldots,x_i,x_i')} = x_1, \quad \ldots, \quad \mathcal{X}_i^{(x_1,\ldots,x_i,x_i')} = x_i,$

$\mathcal{X}_1'^{(x_1,\ldots,x_i,x_i')} = x_1, \quad \ldots, \quad \mathcal{X}_{i-1}'^{(x_1,\ldots,x_i,x_i')} = x_{i-1}, \quad \mathcal{X}_i'^{(x_1,\ldots,x_i,x_i')} = x_i'.$

(ii) $\left(\mathcal{X}_{i+1}^{(x_1,\ldots,x_i,x_i')}, \ldots, \mathcal{X}_\mathcal{N}^{(x_1,\ldots,x_i,x_i')}\right)$

$\sim \mathcal{L}(\mathcal{X}_{i+1}, \ldots, \mathcal{X}_\mathcal{N} | \mathcal{X}_1 = x_1, \ldots, \mathcal{X}_i = x_i),$

$\left(\mathcal{X}'^{(x_1,\ldots,x_i,x_i')}_{i+1}, \ldots, \mathcal{X}'^{(x_1,\ldots,x_i,x_i')}_\mathcal{N}\right)$

$\sim \mathcal{L}(\mathcal{X}_{i+1}, \ldots, \mathcal{X}_\mathcal{N} | \mathcal{X}_1 = x_1, \ldots, \mathcal{X}_{i-1} = x_{i-1}, \mathcal{X}_i = x_i').$

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

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

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

Remark 3.2.2. The definition says that a Marton coupling is a set of couplings between the distributions $\mathcal{L}(\mathcal{X}_{i+1},\ldots,\mathcal{X}_{\mathcal{N}}|\mathcal{X}_1 = x_1,\ldots,\mathcal{X}_i = x_i)$ and $\mathcal{L}(\mathcal{X}_{i+1},\ldots,\mathcal{X}_{\mathcal{N}}|\mathcal{X}_1 = x_1,\ldots,\mathcal{X}_{i-1} = x_{i-1},\mathcal{X}_i = x_i')$ for every $x_1,\ldots,x_i,x_i'$, 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

$$\Gamma_{i,j} \geq \sup_{x_1,\ldots,x_i,x_i'} d_{\mathrm{TV}}\left[\mathcal{L}(\mathcal{X}_j|\mathcal{X}_1 = x_1,\ldots,\mathcal{X}_i = x_i),\right.$$
$$\left.\mathcal{L}(\mathcal{X}_j|\mathcal{X}_1 = x_1,\ldots,\mathcal{X}_{i-1} = x_{i-1},\mathcal{X}_i = x_i')\right],$$

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 $\mathcal{L}(\mathcal{X}_{i+1},\ldots,\mathcal{X}_{\mathcal{N}}|\mathcal{X}_1 = x_1,\ldots,\mathcal{X}_i = x_i)$ and $\mathcal{L}(\mathcal{X}_{i+1},\ldots,\mathcal{X}_{\mathcal{N}}|\mathcal{X}_1 = x_1',\ldots,\mathcal{X}_i = x_i')$. 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 3.2.3 (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 $\hat{X} := (\hat{X}_1, \ldots, \hat{X}_n)$ is a *partition of a vector of random variables* $X = (X_1, \ldots, X_N)$ if $(\hat{X}_i)_{1 \leq i \leq n}$ is a partition of the set $\{X_1, \ldots, X_N\}$. 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 $s(\hat{X}) := \max_{1 \leq i \leq n} s(\hat{X}_i)$ 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 3.2.4 (Marton coupling for Markov chains)**.** *Suppose that $X_1, \ldots, X_N$ 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 $s(\hat{X}) \leq \tau(\epsilon)$, and a Marton coupling for for this*

partition $\hat{X}$ *whose mixing matrix* Γ *satisfies*

$$\Gamma = (\Gamma_{i,j})_{i,j \le n} \le \begin{pmatrix} 1 & 1 & \epsilon & \epsilon^2 & \epsilon^3 & \dots \\ 0 & 1 & 1 & \epsilon & \epsilon^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \tag{3.2.1}$$

with the inequality meant in each element of the matrices.

Remark 3.2.5. 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 Goldstein (1978/79), equation (2.1) on page 482 of Fiebig (1993), and Proposition 2 on page 442 of Samson (2000)).

Proposition 3.2.6 (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* $X = (X_1, \ldots, X_n), Y = (Y_1, \ldots, Y_n)$ *such that* $\mathcal{L}(X) = P$*,* $\mathcal{L}(Y) = Q$*, and* $\mathbb{P}(X_i \neq Y_i) \le d_{\mathrm{TV}}(\mathcal{L}(X_i, \ldots, X_n), \mathcal{L}(Y_i, \ldots, Y_n))$*.*

Remark 3.2.7. Marton (1996b) assumes maximal coupling in each step, corresponding to

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

$$a := \sup_{x,y\in\Omega} d_{\mathrm{TV}}(P(x,\cdot), P(y,\cdot)). \tag{3.2.2}$$

Samson (2000), Chazottes, Collet, Külske, and Redig (2007), Chazottes and Redig (2009), Kontorovich (2007) uses the Marton coupling generated by Proposition 3.2.6. Marton (2003) shows that Marton couplings different from those generated by Proposition 3.2.6 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). Our main contribution is the introduction of the technique of partitioning.

Remark 3.2.8. In the case of time homogeneous Markov chains, Marton couplings (Definition 3.2.1) are in fact equivalent to couplings (X, X') between the distributions $\mathcal{L}(X_1, \ldots, X_N | X_0 = x_0)$ and $\mathcal{L}(X_1, \ldots, X_N | X_0 = x_0')$. Since the seminal paper Doeblin (1938), 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 Rosenthal (1997), the same bound holds if $X_T = X_T'$ (that is, the two chains are equal at a single time).

3.2.2 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 3.2.9 (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*

$$C_i(c) := \sum_{j \in \mathcal{I}(\hat{X}_i)} c_j \text{ for } i \leq n. \tag{3.2.3}$$

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

$$f(x) - f(y) \leq \sum_{i=1}^{n} c_i \mathbb{1}[x_i \neq y_i] \tag{3.2.4}$$

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

$$\log \mathbb{E}\left(e^{\lambda(f(X) - \mathbb{E}f(X))}\right) \leq \frac{\lambda^2 \cdot \|\Gamma \cdot C(c)\|^2}{8} \leq \frac{\lambda^2 \cdot \|\Gamma\|^2 \|c\|^2 s(\hat{X})}{8}. \tag{3.2.5}$$

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

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X)| \geq t\right) \leq 2 \exp\left(\frac{-2t^2}{\|\Gamma \cdot C(c)\|^2}\right), \tag{3.2.6}$$

Remark 3.2.10. Most of the results presented in this chapter are similar to (3.2.6), 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 3.2.11 (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*

$$\tau_{\min} := \inf_{0 \leq \epsilon < 1} \tau(\epsilon) \cdot \left(\frac{2-\epsilon}{1-\epsilon} \right)^2. \tag{3.2.7}$$

Suppose that $f : \Lambda \to \mathbb{R}$ *satisfies* (3.2.4) *for some* $c \in \mathbb{R}_+^N$*. Then for any* $t \geq 0$*,*

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X)| \geq t \right) \leq 2 \exp\left(\frac{-2t^2}{\|c\|^2 \tau_{\min}} \right). \tag{3.2.8}$$

Remark 3.2.12. It is easy to show that for time homogeneous chains,

$$\tau_{\min} \leq \inf_{0 \leq \epsilon < 1} t_{\text{mix}}(\epsilon/2) \cdot \left(\frac{2-\epsilon}{1-\epsilon} \right)^2 \leq 9 t_{\text{mix}}. \tag{3.2.9}$$

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 Lubetzky and Sly (2009)). If this happens, then $\tau_{\min} \approx t_{\text{mix}}$.

Remark 3.2.13. In Example 3.2.17, 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 Gyori and Paulin (2014), 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 Samson (2000) and Marton (2003), 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.

3.2.3 Applications

Example 3.2.14 (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

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

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

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

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

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

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

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

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 3.2.15 (Hidden Markov chains)**.** Let $\tilde{X}_1,\ldots,\tilde{X}_N$ be a Markov chain (not necessarily homogeneous) taking values in $\tilde{\Lambda} = \tilde{\Lambda}_1 \times \ldots \times \tilde{\Lambda}_N$, with distribution $\tilde{P}$. Let $X_1,\ldots,X_N$ be random variables taking values in $\Lambda = \Lambda_1 \times \ldots \times \Lambda_N$ 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 $X_1, \ldots, X_N$ a *hidden Markov chain*.

Concentration inequalities for hidden Markov chains have been investigated in Kontorovich (2006), see also Kontorovich (2007), 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 3.2.16 (McDiarmid's inequality for hidden Markov chains)**.** *Let $\tilde{\tau}(\epsilon)$ denote the mixing time of the underlying chain $\tilde{X}_1, \ldots, \tilde{X}_N$, then Corollary 3.2.11 also applies to hidden Markov chains, with $\tau(\epsilon)$ replaced by $\tilde{\tau}(\epsilon)$ in* (3.2.7)*.*

Proof. It suffices to notice that $(X_1, \tilde{X}_1), (X_2, \tilde{X}_2), \ldots$ 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 (3.2.4) as a function of $X_1, \ldots, X_N$, and it does not depends on $\tilde{X}_1, \ldots, \tilde{X}_N$, 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 3.2.11. □

Example 3.2.17 (Convergence of empirical distribution in total variational distance)**.** Let $X_1, \ldots, X_n$ 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, $d(X_1, \ldots, X_n) := d_{\text{TV}}(\pi_{em}(x), \pi)$.

Proposition 3.2.18. *For any $t \geq 0$,*

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

Proof. The result is an immediate consequence of Corollary 3.2.11, by noticing that the function d satisfies (3.2.4) 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 3.3.16 of Section 3.3, 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 Dvoretzky, Kiefer, and Wolfowitz (1956), Massart (1990)) to the total variational distance case, for Markov chains.

Note that a similar bound was obtained in Kontorovich and Weiss (2012). The main advantage of Proposition 3.2.18 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 Kontorovich and Weiss (2012), where the constants are proportional to a quantity similar to $1/(1-a)^2$ (defined in (3.2.2)).

3.3 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 Lezaud (1998a).

Firstly, in Section 3.3.1, 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 3.3.2, 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 3.3.3, we state propositions that allow us to extend these results to non-stationary chains, and to unbounded functions.

Finally, Section 3.3.4 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 3.3.1. Everywhere in this section, we assume that $X = (X_1, \dots, X_n)$ 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.

3.3.1 Preliminaries

We call a Markov chain $X_1, X_2, \dots$ 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,

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx) \text{ for every } x, y \in \Omega. \tag{3.3.1}$$

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 $\langle f, g\rangle_\pi = \int fg^* \mathrm{d}\pi$, and norm $\|f\|_{2,\pi} := \langle f, f\rangle_\pi^{1/2} = (\mathbb{E}_\pi(f^2))^{1/2}$. 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

$$P^*(x, dy) := \frac{P(y, dx)}{\pi(dx)} \cdot \pi(dy). \tag{3.3.2}$$

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

$$\gamma_{\mathrm{ps}} := \max_{k \geq 1} \left\{ \gamma((\boldsymbol{P}^*)^k \boldsymbol{P}^k)/k \right\}, \tag{3.3.3}$$

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

Remark 3.3.1. The pseudo spectral gap is a generalization of spectral gap of the multiplicative reversiblization ($\gamma(\boldsymbol{P}^*\boldsymbol{P})$), see Fill (1991). We apply it to hypothesis testing for coin tossing (Example 3.3.25). Another application is given in Paulin (2013), 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

$$N_q := \mathbb{E}_\pi \left(\left(\frac{\mathrm{d}\, q}{\mathrm{d}\, \pi} \right)^2 \right) = \int_{x \in \Omega} \frac{\mathrm{d}\, q}{\mathrm{d}\, \pi}(x) q(\mathrm{d}x). \tag{3.3.4}$$

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 3.6.2).

Proposition 3.3.2 (Relation between mixing time and spectral gap)**.** *Suppose that our chain is reversible. For uniformly ergodic chains, for* $0 \leq \epsilon < 1$*,*

$$\gamma^* \geq \frac{1}{1+\tau(\epsilon)/\log(1/\epsilon)}, \text{ in particular, } \gamma^* \geq \frac{1}{1+t_{\text{mix}}/\log(2)}. \tag{3.3.5}$$

For arbitrary initial distribution q*, we have*

$$d_{\text{TV}}\left(q\boldsymbol{P}^n, \pi\right) \leq \frac{1}{2}(1-\gamma^*)^n \cdot \sqrt{N_q - 1}, \tag{3.3.6}$$

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

$$t_{\text{mix}}(\epsilon) \leq \frac{2\log(1/(2\epsilon)) + \log(1/\pi_{\min})}{2\gamma^*}, \text{ in particular,} \tag{3.3.7}$$

$$t_{\text{mix}} \leq \frac{2\log(2) + \log(1/\pi_{\min})}{2\gamma^*}, \tag{3.3.8}$$

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

Proposition 3.3.3 (Relation between mixing time and pseudo spectral gap)**.** *For uniformly ergodic chains, for* $0 \leq \epsilon < 1$*,*

$$\gamma_{\text{ps}} \geq \frac{1-\epsilon}{\tau(\epsilon)}, \text{ in particular, } \gamma_{\text{ps}} \geq \frac{1}{2t_{\text{mix}}}. \tag{3.3.9}$$

For arbitrary initial distribution q*, we have*

$$d_{\text{TV}}\left(q\boldsymbol{P}^n, \pi\right) \leq \frac{1}{2}(1-\gamma_{\text{ps}})^{(n-1/\gamma_{\text{ps}})/2} \cdot \sqrt{N_q - 1}, \tag{3.3.10}$$

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

$$t_{\text{mix}}(\epsilon) \leq \frac{1 + 2\log(1/(2\epsilon)) + \log(1/\pi_{\min})}{\gamma_{\text{ps}}}, \text{ in particular,} \tag{3.3.11}$$

$$t_{\text{mix}} \leq \frac{1 + 2\log(2) + \log(1/\pi_{\min})}{\gamma_{\text{ps}}}. \tag{3.3.12}$$

3.3.2 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 3.6.2). 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 3.3.12 of the next section, we will generalise these bounds to the non-stationary case.

Theorem 3.3.4 (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*

$$\sigma^2_{\text{as}} := \lim_{N\to\infty} N^{-1}\text{Var}_\pi\left(f(X_1) + \ldots + f(X_N)\right). \tag{3.3.13}$$

Then

$$\text{Var}_\pi\left[f(X_1) + \ldots + f(X_n)\right] \leq \frac{2nV_f}{\gamma}, \tag{3.3.14}$$

$$\left|\text{Var}_\pi\left[f(X_1) + \ldots + f(X_n)\right] - n\sigma^2\right| \leq 4V_f/\gamma^2. \tag{3.3.15}$$

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

$$\mathrm{Var}_\pi \left[f_1(X_1) + \ldots + f_n(X_n) \right] \leq \frac{2}{\gamma^*} \sum_{i=1}^{n} \mathrm{Var}_\pi \left[f_i(X_i) \right]. \tag{3.3.16}$$

Remark 3.3.5. 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 (3.3.14). 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 (3.3.16).

Theorem 3.3.6 (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^2_{as} be as in Theorem 3.3.4. Then*

$$\mathrm{Var}_\pi \left[f(X_1) + \ldots + f(X_n) \right] \leq \frac{4nV_f}{\gamma_{\mathrm{ps}}}, \text{ and} \tag{3.3.17}$$

$$\left| \mathrm{Var}_\pi \left[f(X_1) + \ldots + f(X_n) \right] - n\sigma^2_{\mathrm{as}} \right| \leq 16 V_f / \gamma^2_{\mathrm{ps}}. \tag{3.3.18}$$

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

$$\mathrm{Var}_\pi \left[f_1(X_1) + \ldots + f_n(X_n) \right] \leq \frac{4}{\gamma_{\mathrm{ps}}} \sum_{i=1}^{n} \mathrm{Var}_\pi \left[f_i(X_i) \right]. \tag{3.3.19}$$

Theorem 3.3.7 (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^2_{as} be as in Theorem 3.3.4. Let $S := \sum_{i=1}^n f(X_i)$, then*

$$\mathbb{P}_\pi(|S - \mathbb{E}_\pi(S)| \geq t) \leq 2\exp\left(-\frac{t^2}{2n(\sigma^2_{\text{as}} + 0.8V_f) + 10tC/\gamma}\right), \tag{3.3.20}$$

and we also have

$$\mathbb{P}_\pi(|S - \mathbb{E}_\pi(S)| \geq t) \leq 2\exp\left(-\frac{t^2 \cdot \gamma}{4nV_f + 10tC}\right). \tag{3.3.21}$$

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 $V_{S'} := \sum_{i=1}^n \text{Var}_\pi(f_i)$, then

$$\mathbb{P}_\pi(|S' - \mathbb{E}_\pi(S')| \geq t) \leq 2\exp\left(-\frac{t^2 \cdot (2\gamma^* - (\gamma^*)^2)}{8V_{S'} + 20tC}\right), \tag{3.3.22}$$

Remark 3.3.8. The inequality (3.3.20) is an improvement over the earlier result of Lezaud (1998a), because it uses the asymptotic variance σ^2_{as}. In fact, typically $\sigma^2_{\text{as}} \gg V_f$, so the bound roughly equals $2\exp\left(-\frac{t^2}{2n\sigma^2_{\text{as}}}\right)$ for small values of t, which is the best possible given the asymptotic normality of the sum. Note that a result very similar to (3.3.20) has been obtained for continuous time Markov processes by Lezaud (2001).

Theorem 3.3.9 (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 3.3.4. Let

$S := \sum_{i=1}^n f(X_i)$, then

$$\mathbb{P}_\pi(|S - \mathbb{E}_\pi(S)| \geq t) \leq 2\exp\left(-\frac{t^2 \cdot \gamma_{\rm ps}}{8(n + 1/\gamma_{\rm ps})V_f + 20tC}\right). \tag{3.3.23}$$

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 $k_{\rm ps}$ is a the smallest positive integer such that

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

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

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

Then

$$\mathbb{P}_\pi(|S' - \mathbb{E}_\pi(S')| \geq t) \leq 2\exp\left(-\frac{t^2 \cdot \gamma_{\rm ps}}{8V_{S'} + 20tC \cdot M/k_{\rm ps}}\right). \tag{3.3.24}$$

Remark 3.3.10. The bound (3.4.30) is of similar form as (3.3.23) (nV_f is replaced by $V_{S'}$), the main difference is that instead of $20tC$, now we have $20tC \cdot M/k_{\rm ps}$ in the denominator. We are not sure whether the $M/k_{\rm ps}$ term is necessary, or it can be replaced by 1. Note that the bound (3.4.30) also applies if we replace V_i by $V_i' \geq V_i$ for each $1 \leq i \leq n$. In such a way, $M/k_{\rm ps}$ can be decreased, at the cost of increasing $V_{S'}$.

Remark 3.3.11. Theorems 3.3.7 and 3.3.9 can be applied to bound the error of MCMC simulations, see Gyori and Paulin (2014) for more details and examples.

The generalisation to sums of the form $f_1(X_1) + \dots f_n(X_n)$ can be used for "time discounted" sums, see Example 3.3.23.

3.3.3 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 $f_1(X_1) + \dots + f_n(X_n)$, with $X_1, \dots, X_n$ 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 3.6.2.

Proposition 3.3.12 (Bounds for non-stationary chains)**.** *Let $X_1, \dots, X_n$ be a time homogenous Markov chain with state space Ω, and stationary distribution π. Suppose that $g(X_1, \dots, X_n)$ is real valued measurable function. Then*

$$\mathbb{P}_q(g(X_1, \dots, X_n) \geq t) \leq N_q^{1/2} \cdot [\mathbb{P}_\pi(g(X_1, \dots, X_n) \geq t)]^{1/2}, \tag{3.3.25}$$

for any distribution q on Ω (N_q was defined in (3.3.4)). Now suppose that we "burn" the first t_0 observations, and we are interested in bounds on a function h of $X_{t_0+1}, \dots, X_n$. Firstly,

$$\mathbb{P}_q(h(X_{t_0+1}, \dots, X_n) \geq t) \leq N_{q\boldsymbol{P}^{t_0}}^{1/2} \cdot [\mathbb{P}_\pi(h(X_1, \dots, X_n) \geq t)]^{1/2}, \tag{3.3.26}$$

moreover,

$$\mathbb{P}_q(h(X_{t_0+1},\ldots,X_n)\geq t)\leq \mathbb{P}_\pi(h(X_{t_0+1},\ldots,X_n)\geq t)+d_{\mathrm{TV}}\left(q\boldsymbol{P}^{t_0},\pi\right). \tag{3.3.27}$$

Proposition 3.3.13 (Further bounds for non-stationary chains)**.** *In Proposition 3.3.12, $N_{q\boldsymbol{P}^{t_0}}$ can be further bounded. For reversible chains, we have*

$$N_{q\boldsymbol{P}^{t_0}}\leq 1+(N_q-1)\cdot(1-\gamma^*)^{2t_0}, \tag{3.3.28}$$

while for non-reversible chains,

$$N_{q\boldsymbol{P}^{t_0}}\leq 1+(N_q-1)\cdot(1-\gamma_{\mathrm{ps}})^{2(t_0-1/\gamma_{\mathrm{ps}})}. \tag{3.3.29}$$

Similarly, $d_{\mathrm{TV}}\left(q\boldsymbol{P}^n,\pi\right)$ can be further bounded too. For reversible chains, we have, by (3.3.6)*,*

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

For non-reversible chains, by (3.3.10)*,*

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

Finally, for uniformly ergodic Markov chains,

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}^n,\pi\right)\leq\inf_{0\leq\epsilon<1}\epsilon^{\lfloor n/\tau(\epsilon)\rfloor}\leq 2^{-\lfloor n/t_{\mathrm{mix}}\rfloor}. \tag{3.3.30}$$

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 3.3.14 (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 3.3.15. 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, 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.

3.3.4 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 3.3.16 (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 3.2.17, 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 3.3.17. *For stationary, reversible chains,*

$$\mathbb{E}_\pi(d_{\mathrm{TV}}(\pi_{em}, \pi)) \leq \sum_{x\in\Lambda} \min\left(\sqrt{\frac{2\pi(x)}{n\gamma}}, \pi(x)\right). \tag{3.3.31}$$

For stationary, non-reversible chains, (3.3.31) *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 (3.3.14), we have

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

By Jensen's inequality, we obtain that

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

and the statement follows by summing up. The proof of the non-reversible case is

similar, using (3.3.17) to bound the variance. □

It is easy to see that for any stationary distribution π, our bound (3.3.31) 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

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

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

Example 3.3.18 (Estimation of the asymptotic variance)**.** Now we propose an estimator to the asymptotic variance σ^2. For some integer $k \in [1, \hat{N} - \hat{t}_0 - 1]$, let

$$\hat{\sigma}^2(k) := \left(\hat{\gamma}_0 + 2\sum_{i=1}^{k} \hat{\gamma}_i\right) \cdot \frac{\hat{N} - \hat{t}_0 - k}{\hat{N} - \hat{t}_0 - 3k - 1}, \tag{3.3.32}$$

with

$$\hat{\gamma}_i := \frac{\sum_{j=\hat{t}_0+1}^{\hat{N}-k} f(X_j)f(X_{j+i})}{\hat{N} - \hat{t}_0 - k} - \frac{1}{2}\left(\frac{\sum_{j=\hat{t}_0+1}^{\hat{N}-k} f(X_j)}{\hat{N} - \hat{t}_0 - k}\right)^2 - \frac{1}{2}\left(\frac{\sum_{j=\hat{t}_0+i}^{\hat{N}-k+i} f(X_j)}{\hat{N} - \hat{t}_0 - k}\right)^2. \tag{3.3.33}$$

The following two propositions bounds on the bias of $\hat{\sigma}^2(k)$, and state a non-asymptotic error bound for it.

Proposition 3.3.19 (Bias of $\hat{\sigma}^2(k)$)**.** *For stationary, reversible chains, when k is even, the expected value of $\hat{\sigma}^2(k)$ satisfies the following inequality:*

$$-L_k \leq \sigma^2 - \mathbb{E}_\pi(\hat{\sigma}^2(k)) \leq U_k, \tag{3.3.34}$$

with

$$L_k := \left(\min\left(V_f, \frac{2V_f}{\gamma}(1-\gamma^*)^{k+1} \right) + \frac{4V_f}{\gamma^2} \frac{2k+1}{(\hat{N}-\hat{t}_0-k)^2} \right) \cdot \frac{\hat{N}-\hat{t}_0-k}{\hat{N}-\hat{t}_0-3k-1}, \text{ and}$$

$$U_k := \left(\frac{2V_f}{\gamma}(1-\min(\gamma,1))^{k+1} + \frac{4V_f}{\gamma^2} \frac{2k+1}{(\hat{N}-\hat{t}_0-k)^2} \right) \cdot \frac{\hat{N}-\hat{t}_0-k}{\hat{N}-\hat{t}_0-3k-1}.$$

For stationary non-reversible chains, for any $k \geq 1$,

$$|\mathbb{E}_\pi(\hat{\sigma}^2(k)) - \sigma^2| \leq W_k, \tag{3.3.35}$$

with

$$W_k := \frac{4V_f}{\gamma_{\text{ps}}}(1-\gamma_{\text{ps}})^{(k+1-1/\gamma_{\text{ps}})/2} + \frac{16V_f}{\gamma_{\text{ps}}^2} \frac{2k+1}{(\hat{N}-\hat{t}_0-k)^2}.$$

Proposition 3.3.20 (Concentration of $\hat{\sigma}^2(k)$)**.** *Suppose that $f : \Omega \to \mathbb{R}$ satisfies that $\sup_{x\in\Omega} |f(x) - \mathbb{E}_\pi f| \leq C$ for some finite C. In the case of stationary, uniformly ergodic chains, we have for any $t \geq 0$,*

$$\mathbb{P}_\pi(|\hat{\sigma}^2(k) - \mathbb{E}_\pi(\hat{\sigma}^2(k))| \geq t) \leq 2\exp\left(\frac{-t^2(\hat{N}-\hat{t}_0-3k-1)}{288(2k+1)^2 C^4 t_{\text{mix}}} \right), \tag{3.3.36}$$

This implies that for uniformly ergodic reversible chains, with arbitrary initial distribution q, for even $k \geq 2$, any $t \geq 0$,

$$\mathbb{P}_q\left(\sigma^2 - \hat{\sigma}^2(k) \geq U_k + t\right) \leq \exp\left(\frac{-t^2(\hat{N}-\hat{t}_0-3k-1)}{288(2k+1)^2 C^4 t_{\text{mix}}} \right) + E(\hat{t}_0), \tag{3.3.37}$$

and for uniformly ergodic non-reversible chains, for any $k \geq 1$, $t \geq 0$,

$$\mathbb{P}_q\left(\sigma^2 - \hat{\sigma}^2(k) \geq W_k + t\right) \leq \exp\left(\frac{-t^2(\hat{N} - \hat{t}_0 - 3k - 1)}{288(2k+1)^2 C^4 t_{\text{mix}}}\right) + E(\hat{t}_0). \tag{3.3.38}$$

Remark 3.3.21. It is clear that if we increase k, the bias $|\sigma^2 - \mathbb{E}_\pi(\hat{\sigma}^2(k))|$ becomes smaller, but the concentration bounds become weaker.

With the choice

$$\hat{t}_0 := \lfloor 0.1\hat{N} \rfloor, \;\; k := 10 \cdot \left\lfloor \hat{N}^{1/3} \right\rfloor, \;\; \hat{\sigma}^2 := \hat{\sigma}^2(k), \tag{3.3.39}$$

our bounds imply that for bounded functions, $\hat{\sigma}^2$ will be a consistent estimate of σ^2 as $\hat{N} \to \infty$, for any uniformly ergodic Markov chain, irrespectively of the value of the mixing time. In practice, we suggest choosing $\hat{N}$ to be at least 10^6, or higher. Note that via Proposition 3.3.14, the error bound of Proposition 3.3.20 can also be extended to unbounded functions.

We will use the following lemma for the proof of our propositions.

Lemma 3.3.22. *For $t \in \mathbb{N}$, let $\gamma_t := \mathbb{E}_\pi[(f(X_1) - \mathbb{E}_\pi f)(f(X_{t+1}) - \mathbb{E}_\pi f)]$. Then for reversible chains, for $k \geq 2$ even,*

$$-\min\left(\frac{V_f}{2}, \frac{2V_f}{\gamma} \cdot (1-\gamma^*)^{k+1}\right) \leq \sigma^2 - \left(\gamma_0 + 2\sum_{t=1}^{k} \gamma_t\right) \leq \frac{2V_f}{\gamma} \cdot (1 - \min(\gamma, 1))^{k+1}. \tag{3.3.40}$$

For non-reversible chains, we have, for $k \geq 1$,

$$\left|\sigma^2 - \left(\gamma_0 + 2\sum_{t=1}^{k} \gamma_t\right)\right| \leq \frac{4V_f}{\gamma_{\text{ps}}} \cdot (1 - \gamma_{\text{ps}})^{(k+1-1/\gamma_{\text{ps}})/2}. \tag{3.3.41}$$

Proof. Without loss of generality, assume that $\mathbb{E}_\pi f = 0$. Define the operator $\boldsymbol{\pi}$ on $L^2(\pi)$ as $\boldsymbol{\pi}(g)(x) := \mathbb{E}_\pi(g)$. We have $\sigma^2 = \gamma_0 + 2\sum_{t=1}^{\infty}\gamma_t$, thus

$$\sigma^2 - \left(\gamma_0 + 2\sum_{t=1}^{k}\gamma_t\right) = 2\sum_{t=k+1}\gamma_t = 2\left\langle f, \left(\sum_{t=k+1}^{\infty}\boldsymbol{P}^t\right) f\right\rangle_\pi$$
$$= 2\left\langle f, \left(\sum_{t=k+1}^{\infty}(\boldsymbol{P}-\boldsymbol{\pi})^t\right) f\right\rangle_\pi = 2\left\langle f, (\boldsymbol{P}-\boldsymbol{\pi})^{k+1}\left(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi})\right)^{-1} f\right\rangle_\pi.$$

For reversible chains, one one hand, we can write $\|\boldsymbol{P}-\boldsymbol{\pi}\|_{2,\pi} \le 1-\gamma^*$, and

$$\|\left(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi})\right)^{-1}\|_{2,\pi} = 1/\gamma,$$

thus

$$\left|\sigma^2 - (\gamma_0 + 2\sum_{t=1}^{k}\gamma_t)\right| \le \frac{2V_f}{\gamma}\cdot(1-\gamma^*)^{k+1}. \tag{3.3.42}$$

On the other hand, we can express the self-adjoint operator $(\boldsymbol{P}-\boldsymbol{\pi})^{k+1}\left(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi})\right)$ as a sum of positive and negative parts (we also use the fact that $k+1$ is odd):

$$(\boldsymbol{P}-\boldsymbol{\pi})^{k+1}\left(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi})\right)^{-1} = \left((\boldsymbol{P}-\boldsymbol{\pi})_+^{k+1} - (\boldsymbol{P}-\boldsymbol{\pi})_-^{k+1}\right)\left(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi})\right)^{-1}.$$

Now it is easy to see that

$$\|\left(\boldsymbol{P}-\boldsymbol{\pi}\right)_+^{k+1}\left(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi})\right)^{-1}\|_{2,\pi} \le \min(\gamma,1)^{k+1}/\gamma, \text{ and}$$
$$\|\left(\boldsymbol{P}-\boldsymbol{\pi}\right)_-^{k+1}\left(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi})\right)^{-1}\|_{2,\pi} \le 1/2,$$

thus

$$
\begin{aligned}
&-\min\left(V_f, \frac{2V_f}{\gamma}(1-\gamma^*)^{k+1}\right) \leq 2\left\langle f, (\boldsymbol{P}-\boldsymbol{\pi})^{k+1}\left(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi})\right)^{-1} f\right\rangle_{\pi} \\
&\leq \frac{2V_f}{\gamma}\left(1-\min(\gamma,1)\right)^{k+1}.
\end{aligned}
$$

Combining this and (3.3.42) leads to (3.3.40). For non-reversible chains, by the proof of Theorem 3.3.6, we have that $\| \left(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi})\right)^{-1} \|_{2,\pi} \leq 2/\gamma_{\mathrm{ps}}$, and $\| \left(\boldsymbol{P}-\boldsymbol{\pi}\right)^{k+1} \|_{2,\pi} \leq (1-\gamma_{\mathrm{ps}})^{(k+1-1/\gamma_{\mathrm{ps}})/2}$, thus (3.3.41) follows. □

Now we turn to the proof of the two propositions. First we prove the expectation bounds, and then the concentration bounds.

Proof of Proposition 3.3.19. For reversible chains, for $0 \leq i \leq k$, from Chebyshev's inequality (Theorem 3.3.4), we obtain

$$
\begin{aligned}
&\left|\mathbb{E}_{\pi}\left(\frac{1}{2}\left(\frac{\sum_{j=t_0+1}^{\hat{N}-k} f(X_j)}{\hat{N}-\hat{t}_0-k}\right)^2 + \frac{1}{2}\left(\frac{\sum_{j=t_0+i}^{\hat{N}-k+i} f(X_j)}{\hat{N}-\hat{t}_0-k}\right)^2\right) - \left[(\mathbb{E}_{\pi} f)^2 + \frac{\sigma^2}{\hat{N}-\hat{t}_0-k}\right]\right| \\
&\leq \frac{4V_f}{\gamma^2}\cdot\frac{1}{(\hat{N}-\hat{t}_0-k)^2},
\end{aligned}
$$

and thus it follows that

$$
\left|\mathbb{E}_{\pi}(\hat{\gamma}_i) - \left(\gamma_i - \frac{\sigma^2}{\hat{N}-\hat{t}_0-k}\right)\right| \leq \frac{4V_f}{\gamma^2}\cdot\frac{1}{(\hat{N}-\hat{t}_0-k)^2}.
$$

Summing up in i, and using (3.3.40) leads to

$$
\begin{aligned}
&- K_f - \min\left(V_f, \frac{2V_f}{\gamma}(1-\gamma^*)^{k+1}\right) \leq \sigma^2 - \left(\hat{\gamma}_0 + 2\sum_{i=1}^{k}\hat{\gamma}_i + \frac{\sigma^2(2k+1)}{\hat{N}-\hat{t}_0-k}\right) \\
&\leq K_f + \frac{2V_f}{\gamma}\cdot(1-\min(\gamma,1))^{k+1},
\end{aligned}
$$

where $K_f := \frac{4V_f}{\gamma^2}\cdot\frac{(2k+1)}{(\hat{N}-\hat{t}_0-k)^2}$. Now putting together the terms involving σ^2, and dividing by $\frac{\hat{N}-\hat{t}_0-3k-1}{\hat{N}-\hat{t}_0-k}$ leads to (3.3.34). The proof of (3.3.35) is similar. □

Proof of Proposition 3.3.20. Firstly, it is easy to show for any $0 \leq i \leq k$, $\hat{\gamma}_i$ does not change if we replace the function f by $f - \mathbb{E}_\pi f$, thus $\sigma^2(k)$ remains the same under such transformation. Now a simple computation shows that changing the value of X_j, for $\hat{t}_0 + 1 \leq j \leq \hat{N}$, can only change $\hat{\gamma}_i$ at most by $8C^2/(\hat{N}-\hat{t}_0-k)$, and thus it can only change the value of $\hat{\sigma}^2(k)$ at most by $8(2k+1)C^2/(\hat{N}-\hat{t}_0-3k-1)$. From this (the so called Hamming-Lipschitz property), using McDiarmid's bounded differences inequality for Markov chains (Corollary 3.2.11), we can deduce (3.3.36). Finally, (3.3.37) and (3.3.38) follow by combining this with the bounds on the bias. □

Example 3.3.23 (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 $X_1, X_2, \ldots$ 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

$$W := \sum_{i=1}^{\infty} X_i (1+r)^{-i}. \tag{3.3.43}$$

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 3.3.7) with $f_i(X_i) = X_i(1+r)^{-i}$ and $C = 1$, and use a limiting argument, to obtain that

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

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 3.3.9).

Example 3.3.24 (Hypothesis testing)**.** The following example was inspired by Hu (2011). Suppose that we have a sample $X = (X_1, X_2, \ldots, X_n)$ 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 $\xi \in \mathbb{R}$,

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

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 $i \geq 1$. Then $(Y_i)_{i\geq 1}$ 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

$$\hat{T}(Y) := \sum_{i=1}^{n-1} \log\left(\frac{P_0(Y_i)}{P_1(Y_i)}\right) = \sum_{i=1}^{n-1} \log\left(\frac{P_0(X_i, X_{i+1})}{P_1(X_i, X_{i+1})}\right), \tag{3.3.44}$$

then

$$\frac{T(X)}{n-1} = \frac{\log(\pi_0(X_1)/\pi_1(X_1))}{n-1} + \frac{\hat{T}(Y)}{n-1}. \tag{3.3.45}$$

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 $\left|\frac{\log(\pi_0(X_1)/\pi_1(X_1))}{n-1}\right| \leq \frac{\delta}{n-1}$, implying that $|T(X)/(n-1) - \hat{T}(Y)/(n-1)| \leq \delta/(n-1)$. Moreover, we also have $|\log P_0(Y_i) - \log P_1(Y_i)| \leq \delta$.

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

$$\mathbb{P}_0\left(\frac{T(X)}{n-1} \leq \xi\right) \leq \exp\left(-\frac{(J_0 - \delta/(n-1) - \xi)^2 (n-1)\gamma_{\mathrm{ps}}(\boldsymbol{Q}_0)}{8V_0 + 20\delta\cdot(J_0 - \delta/(n-1) - \xi)}\right), \tag{3.3.46}$$

$$\mathbb{P}_1\left(\frac{T(X)}{n-1} \geq \xi\right) \leq \exp\left(-\frac{(\xi - J_1 - \delta/(n-1))^2 (n-1)\gamma_{\mathrm{ps}}(\boldsymbol{Q}_1)}{8V_1 + 20\delta\cdot(\xi - J_1 - \delta/(n-1))}\right). \tag{3.3.47}$$

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 3.3.25 (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 Diaconis, Holmes, and Montgomery (2007), 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. Using Example 3.3.24, 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 3.3.24, 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 (3.3.2),

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

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 $\gamma((\boldsymbol{Q}_1^*)^2\boldsymbol{Q}_1^2) = 0.96$, thus $\gamma_{\text{ps}}(\boldsymbol{Q}_0) = 0.5$, $\gamma_{\text{ps}}(\boldsymbol{Q}_1) = 0.48$. 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 (3.3.46) and (3.3.47), and choosing $\xi = 0$, we obtain the following error bounds.

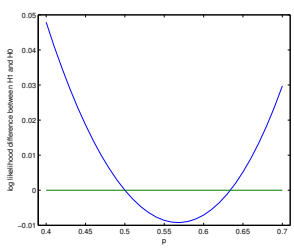
$$\text{Type-I error.} \quad \mathbb{P}_0(T(X)/(n-1) \le \xi) \le \exp(-4.120) = 0.0150, \tag{3.3.48}$$

$$\text{Type-II error.} \quad \mathbb{P}_1(T(X)/(n-1) \ge \xi) \le \exp(-4.133) = 0.0160. \tag{3.3.49}$$

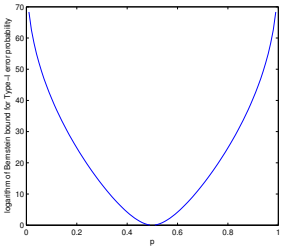
The actual value of $T(X)/(n-1)$ on our data is $\tilde{T}/(n-1) = -7.080 \cdot 10^{-3}$. Since $\tilde{T}/(n-1) < \xi$, we reject H0 (Bernoulli i.i.d. trials).

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

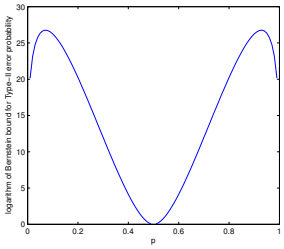
Figure 3.1: Hypothesis testing for different values of the parameter p



(a) Log-likelihood difference



(b) Logarithm of bound on Type-I error



(c) Logarithm of bound on Type-II error

outside of this interval, we stand by H0. Three plots in Figure 3.1 show the log-likelihood differences, and 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 $\hat{p} \approx 0.57$.

3.4 Continuous time Markov processes

In this section, we are going to generalise our previous results for Markov chains to continuous time Markov processes.

Firstly, in Section 3.4.1, we introduce Markov processes, and define the continuous time versions of our previous notions, including the mixing time, and the spectral gap. We also state propositions concerning the relations between mixing time and spectral gap, the way to get bounds for non-stationary processes from bounds for stationary ones, and the use of truncation to handle unbounded functions.

After this, in Section 3.4.2 we state our results for Markov processes: variance bounds and Bernstein-type inequalities for integrals of the form $\int_{t=0}^{T} f_t(X_t)\mathrm{d}t$, and a version of McDiarmid's bounded differences inequality for uniformly ergodic Markov processes.

Finally, Section 3.4.4 contains two applications, the average number of persons waiting in an M/M/1 queue, and the total variational distance of the empirical measure to the stationary distribution of Markov processes.

3.4.1 Preliminaries

In this section, we will first define Markov processes, then generalize the notions of Sections 3.1.1 and 3.3.1 to them.

We call a collection of random variables $\{X_t\}_{t\geq 0}$ (taking values in Polish spaces $\{\Lambda_t\}_{t\geq 0}$, and defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$) a Markov process if they satisfy the Markov property: for every $0 \leq s < t$, for every $y \in \Lambda_t$,

$$\mathbb{P}(X_t \in \mathrm{d}y | \{X_r\}_{0\leq r\leq s}) = \mathbb{P}(X_t \in \mathrm{d}y | X_s). \tag{3.4.1}$$

In this case, we define, for $0 \leq s < t$,

$$P_{s,t}(x, \mathrm{d}y) := \mathbb{P}(X_t \in \mathrm{d}y | X_s = x) \tag{3.4.2}$$

the *Markov kernel* of the time inhomogeneous Markov process $\{X_t\}_{t\geq 0}$. By the fact that $\{X_t\}_{t\geq 0}$ are defined on the same probability space, such a Markov kernel always satisfies the Chapman-Kolmogorov equations: for any $0 \leq s < r < t$,

$$P_{s,t}(x, \mathrm{d}y) = \int_{z \in \Lambda_r} P_{s,r}(x, \mathrm{d}z) P_{r,t}(z, \mathrm{d}y). \tag{3.4.3}$$

We say that $\{X_t\}_{t \geq 0}$ is a *time homogenous Markov process* if the state space Λ_t is the same, Ω, for every $t \geq 0$, and $P_{s,t}(x, \mathrm{d}y) = P_{0,t-s}(x, \mathrm{d}y)$ for every $x, y \in \Omega$, $0 \leq s < t$. In this case, we define $P_t(x, \mathrm{d}y) := P_{0,t-s}(x, \mathrm{d}y)$.

For time homogeneous Markov processes, we say that a distribution π on Ω is a *stationary distribution* if for every $x \in \Lambda$, every $t > 0$,

$$\int_{x \in \Omega} P_t(x, \mathrm{d}y) \pi(\mathrm{d}x) = \pi(\mathrm{d}y). \tag{3.4.4}$$

Now we define uniform and geometric ergodicity for Markov processes:

Definition 3.4.1. A time homogeneous Markov process $\{X_t\}_{t \geq 0}$, with stationary distribution π, state space Ω, and transition kernels $\{P_t(x, dy)\}_{t>0}$ is *uniformly ergodic* if

$$\sup_{x \in \Lambda} d_{\mathrm{TV}} \left(P_t(x, \cdot), \pi \right) \leq M \rho^t, \text{ for every } t > 0,$$

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

$$d_{\mathrm{TV}} \left(P_t(x, \cdot), \pi \right) \leq M(x) \rho^t, \text{ for every } t > 0$$

for some $\rho < 1$, where $M(x) < \infty$ for π a.e. $x \in \Omega$.

Now we are going to define mixing times for time homogeneous and inhomogeneous Markov processes.

Definition 3.4.2 (Mixing time for time homogeneous processes)**.** Let $\{X_t\}_{t\geq 0}$ be a time homogeneous Markov process with transition kernels $\{P_t(x,dy)\}_{t>0}$, state space Ω (a Polish space), and stationary distribution π. For $t>0$, let $d^{\text{cont}}(t) := \sup_{x\in\Omega} d_{\text{TV}}\left(P_t(x,\cdot),\pi\right)$, and let

$$t_{\text{mix}}^{\text{cont}}(\epsilon) := \min\{t>0 : d(t)\leq\epsilon\} \text{ and } t_{\text{mix}}^{\text{cont}} := t_{\text{mix}}(1/4).$$

One can easily prove that 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 Markov process.

Definition 3.4.3 (Mixing time for time inhomogeneous processes)**.** Let $\{X_t\}_{t\geq 0}$ be a Markov process with transition kernels $\{P_{s,t}(x,dy)\}_{0\leq s<t}$, state space $\{\Lambda_t\}_{t\geq 0}$ (a Polish space). Let

$$\begin{aligned}\overline{d}^{\text{cont}}(t) &:= \sup_{s:s\geq 0}\sup_{x,y\in\Omega} d_{\text{TV}}\left(P_{s,s+t}(x,\cdot),P_{s,s+t}(y,\cdot)\right),\\ \tau^{\text{cont}}(\epsilon) &:= \inf\left\{t : t>0, \overline{d}^{\text{cont}}(t)\leq\epsilon\right\}.\end{aligned}$$

Remark 3.4.4. One can easily see that in the case of time homogeneous Markov processes, by triangle inequality, one has

$$\tau^{\text{cont}}(2\epsilon)\leq t_{\text{mix}}(\epsilon)\leq\tau^{\text{cont}}(\epsilon). \tag{3.4.5}$$

One can show that $\overline{d}^{\text{cont}}(t)$ is subadditive, i.e.

$$\overline{d}^{\text{cont}}(t+s)\leq\overline{d}^{\text{cont}}(t)+\overline{d}^{\text{cont}}(s), \tag{3.4.6}$$

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

$$\tau^{\text{cont}}(\epsilon^k) \leq k\tau^{\text{cont}}(\epsilon), \text{ and thus } t_{\text{mix}}^{\text{cont}}\left((2\epsilon)^k\right) \leq k t_{\text{mix}}^{\text{cont}}(\epsilon). \tag{3.4.7}$$

Now, we are going to generalise some of the spectral properties from Section 3.3.1 to time homogeneous Markov processes. First of all, we define $L^2(\pi)$, and the operator $\boldsymbol{P}$ corresponding to a Markov kernel P the same way as in Section 3.3.1. We say that the time homogeneous Markov process $\{X_t\}_{t\geq 0}$ is *reversible* with respect to stationary distribution π if for every $t > 0$, the Markov kernel $P_t(x, \mathrm{d}y)$ is reversible (i.e. $P_t(x, \mathrm{d}y)\pi(\mathrm{d}x) = P_t(y, \mathrm{d}x)\pi(\mathrm{d}y)$). Equivalently, this means that $\boldsymbol{P}_t$ is self-adjoint for every $t > 0$.

A family of operators $\{\boldsymbol{P}\}_{t\geq 0}$ is a *stochastic semigroup* if

1. $\boldsymbol{P}_0 = \boldsymbol{I}$,

2. each element $\boldsymbol{P}_t$ is generated by a Markov kernel $P_t(x, \mathrm{d}y)$, and

3. it satisfies the Chapman-Kolmogorov equations: $\boldsymbol{P}_{s+t} = \boldsymbol{P}_s\boldsymbol{P}_t$, $t \geq 0$.

The stochastic semigroup $\{\boldsymbol{P}_t\}_{t\geq 0}$ is *standard* if $\boldsymbol{P}_t \to \boldsymbol{I}$ as $t \downarrow 0$, i.e. for every $f \in L^2(\pi)$, $\lim_{t\downarrow 0}(\boldsymbol{P}_t f)(x) = f(x)$, π - a.s. in x.

For a time homogeneous Markov process $\{X_t\}_{t\geq 0}$ with standard stochastic semigroup $\{\boldsymbol{P}_t\}_{t\geq 0}$, we define its *generator* $\boldsymbol{L}$ as a linear operator from $\mathcal{D}_2(\boldsymbol{L})$ to $\mathcal{D}_2(\boldsymbol{L})$, defined as

$$\boldsymbol{L}(f) := \lim_{t\downarrow 0} \frac{\boldsymbol{P}_t f - f}{t}, \tag{3.4.8}$$

with the domain $\mathcal{D}_2(\boldsymbol{L})$ being the subset of $L^2(\pi)$ such that this limit exists in the $L^2(\pi)$ sense. The Hille-Yosida theory (Yosida (1980)) shows that $\mathcal{D}_2(\boldsymbol{L})$ is a dense

subspace of $L^2(\pi)$. Notice that if the process is reversible, then $\boldsymbol{L}$ is self-adjoint.

We define *spectrum of the process* as

$$S_2 := \Big\{ \lambda \in \mathbb{R} \setminus 0 : (\lambda \boldsymbol{I} - (\boldsymbol{L} + \boldsymbol{L}^*)/2)^{-1} \text{ does not exist as a bounded linear operator on } \mathcal{D}_2(\boldsymbol{L}) \Big\},$$

and the *spectral gap of the process* as

$$\gamma^{\text{cont}} := -\sup\{\lambda : \lambda \in S_2, \lambda \neq 0\} \quad \text{if eigenvalue 0 of } (\boldsymbol{L} + \boldsymbol{L}^*)/2 \text{ is simple,}$$
$$\gamma^{\text{cont}} := 0 \quad \text{otherwise.}$$

Notice that $(\boldsymbol{L} + \boldsymbol{L}^*)/2$ is negative semidefinite, thus $S_2 \subset \mathbb{R}_-$ and $\gamma^{\text{cont}} \geq 0$.

For a time homogeneous Markov process $\{X_t\}_{t\geq 0}$, we define the *pseudo spectral gap of the process* as

$$\gamma_{\text{ps}}^{\text{cont}} := \sup_{t>0} \left\{ \gamma(\boldsymbol{P}_t^* \boldsymbol{P}_t)/t \right\}, \tag{3.4.9}$$

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

The relations between the mixing and spectral properties of time homogeneous Markov processes are given by the following two propositions proposition (which are similar to Propositions 3.3.2 and 3.3.3).

Proposition 3.4.5 (Relation between mixing time and spectral gap for Markov processes)**.** *Suppose that we have a reversible Markov process. For uniformly ergodic Markov processes, for* $0 \leq \epsilon < 1$*,*

$$\gamma^{\text{cont}} \geq \frac{\log(1/\epsilon)}{\tau^{\text{cont}}(\epsilon)}, \text{ in particular, } \gamma^{\text{cont}} \geq \frac{\log(2)}{t_{\text{mix}}^{\text{cont}}}. \tag{3.4.10}$$

For arbitrary initial distribution q, we have for every $t > 0$,

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}_t, \pi\right) \leq \frac{1}{2}(1-\gamma^{\mathrm{cont}})^t \cdot \sqrt{N_q - 1}, \tag{3.4.11}$$

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

$$t_{\mathrm{mix}}^{\mathrm{cont}}(\epsilon) \leq \frac{2\log(1/(2\epsilon)) + \log(1/\pi_{\min})}{2\gamma^{\mathrm{cont}}}, \text{ in particular, } t_{\mathrm{mix}}^{\mathrm{cont}} \leq \frac{2\log(2) + \log(1/\pi_{\min})}{2\gamma^{\mathrm{cont}}} \tag{3.4.12}$$

with N_q and $\pi_{\min}$ defined as in Proposition 3.3.2.

Proposition 3.4.6 (Relation between mixing time and pseudo spectral gap for Markov processes)**.** *For uniformly ergodic Markov processes, for $0 \leq \epsilon < 1$,*

$$\gamma_{\mathrm{ps}}^{\mathrm{cont}} \geq \frac{1-\epsilon}{\tau^{\mathrm{cont}}(\epsilon)}, \text{ in particular, } \gamma_{\mathrm{ps}}^{\mathrm{cont}} \geq \frac{1}{2t_{\mathrm{mix}}^{\mathrm{cont}}}. \tag{3.4.13}$$

When starting from initial distribution q, we have for every $t > 0$,

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}_t, \pi\right) \leq \frac{1}{2}(1-\gamma_{\mathrm{ps}}^{\mathrm{cont}})^{(t-1/\gamma_{\mathrm{ps}}^{\mathrm{cont}})/2} \cdot \sqrt{N_q - 1}, \tag{3.4.14}$$

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

$$t_{\mathrm{mix}}^{\mathrm{cont}}(\epsilon) \leq \frac{1 + 2\log(1/(2\epsilon)) + \log(1/\pi_{\min})}{\gamma_{\mathrm{ps}}^{\mathrm{cont}}}, \text{ in particular,} \tag{3.4.15}$$

$$t_{\mathrm{mix}} \leq \frac{1 + 2\log(2) + \log(1/\pi_{\min})}{\gamma_{\mathrm{ps}}^{\mathrm{cont}}}. \tag{3.4.16}$$

Some of the definitions above were adapted from the survey Bakry (2006). Other good references on the subject are Saloff-Coste (1997), Montenegro and Tetali (2006),

and Wang (2006) (however, in some of these, the authors restrict themselves to heat kernels, a special case of continuous time Markov processes with generator of the form $\boldsymbol{L} = \boldsymbol{P} - \boldsymbol{I}$ for some Markov kernel P).

3.4.2 Results

In this section, we are going to state variance bounds, Bernstein-type inequalities for integrals of the form $\int_{t=0}^{T} f_t(X_t)\mathrm{d}t$, with $(X_t)_{0\leq t\leq T}$ being a time homogeneous Markov process. These bounds will be stated for stationary Markov processes, however, they can be generalised to non-stationary processes by Proposition 3.4.16. We also state a version of McDiarmid's bounded differences inequality for uniformly ergodic Markov processes.

Firstly, we state the variance bounds for reversible and non-reversible processes.

Theorem 3.4.7 (Variance bound for reversible processes)**.** *Let $\{X_t\}_{t\geq 0}$ be a time homogeneous, stationary, reversible Markov process, with distribution $\mathbb{P}$, standard stochastic semigroup $\{\boldsymbol{P}_t\}_{t\geq 0}$, stationary distribution π, and spectral gap γ^{cont}. For $f \in L^2(\pi)$, we define its* variance *as $V_f := \mathrm{Var}_\pi(f)$, and its* asymptotic variance *as*

$$\sigma^2_{\mathrm{cont}} := \lim_{T\to\infty} (1/T) \cdot \mathrm{Var}_\pi \left(\int_{t=0}^{T} f(X_t)\mathrm{d}t \right). \tag{3.4.17}$$

Assume that f satisfies

$$\lim_{N\to\infty} \frac{T}{N} \sum_{k=1}^{N} f\left(X_{\frac{T}{N}\cdot k}\right) = \int_{t=0}^{T} f(X_t)\mathrm{d}t \ \ \mathbb{P} \text{ almost surely}, \tag{3.4.18}$$

$$\lim_{N\to\infty} \frac{T}{N} \sum_{k=1}^{N} \mathbb{E}_\pi\left(f^2\right) = \int_{t=0}^{T} \mathbb{E}_\pi(f^2)\mathrm{d}t. \tag{3.4.19}$$

Then

$$\mathrm{Var}_\pi\left[\int_{t=0}^{T} f(X_t)\mathrm{d}t\right] \leq \frac{2TV_f}{\gamma^{\mathrm{cont}}}, \tag{3.4.20}$$

$$\left|\mathrm{Var}_\pi\left[\int_{t=0}^{T} f(X_t)\mathrm{d}t\right] - T\sigma^2_{\mathrm{cont}}\right| \leq \frac{4V_f}{(\gamma^{\mathrm{cont}})^2}. \tag{3.4.21}$$

Remark. We apply this theorem to M/M/1 queues in Example 3.4.22.

The following theorem generalises this to integrals of the form $\int_{t=0}^{T} f_t(X_t)$.

Theorem 3.4.8. *Let $\{X_t\}_{t\geq 0}$ be as in Theorem 3.4.7. Let $\{f_t\}_{t\geq 0}$ be functions in $L^2(\pi)$ satisfying that $\mathbb{E}_\pi f_t = 0$ for every $t \geq 0$. Assume that*

$$\lim_{N\to\infty} \frac{T}{N}\sum_{k=1}^{N} f_{\frac{T}{N}\cdot k}\left(X_{\frac{T}{N}\cdot k}\right) = \int_{t=0}^{T} f_t(X_t)\mathrm{d}t \ \ \mathbb{P} \text{ almost surely}, \tag{3.4.22}$$

$$\lim_{N\to\infty} \frac{T}{N}\sum_{k=1}^{N} \mathbb{E}_\pi\left(f^2_{\frac{T}{N}\cdot k}\right) = \int_{t=0}^{T} \mathbb{E}_\pi(f_t^2)\mathrm{d}t. \tag{3.4.23}$$

Then

$$\mathrm{Var}_\pi\left[\int_{t=0}^{T} f_t(X_t)\mathrm{d}t\right] \leq \frac{2\int_{t=0}^{T}\mathrm{Var}_\pi\left(f_t\right)\mathrm{d}t}{\gamma^{\mathrm{cont}}}. \tag{3.4.24}$$

Theorem 3.4.9 (Variance bounds for non-reversible Markov processes)**.** *Let $\{X_t\}_{t\geq 0}$ be a time homogeneous Markov process, with distribution $\mathbb{P}$, stationary distribution π, and pseudo spectral gap $\gamma^{\mathrm{cont}}_{\mathrm{ps}}$.*

Let f be a function in $L^2(\pi)$ satisfying the regularity conditions (3.4.18) *and*

(3.4.19). *Let V_f and σ^2_{cont} be as in Theorem 3.4.7. Then*

$$\text{Var}_\pi\left[\int_{t=0}^{T} f(X_t)\text{d}t\right] \leq \frac{4TV_f}{\gamma_{\text{ps}}^{\text{cont}}}, \tag{3.4.25}$$

$$\left|\text{Var}_\pi\left[\int_{t=0}^{T} f(X_t)\text{d}t\right] - T\sigma^2_{\text{cont}}\right| \leq \frac{16V_f}{(\gamma_{\text{ps}}^{\text{cont}})^2}. \tag{3.4.26}$$

More generally, let $\{f_t\}_{t\geq 0}$ be functions in $L^2(\pi)$ satisfying the regularity conditions (3.4.22) *and* (3.4.23) *(essentially Riemann integrability), then*

$$\mathbb{E}_\pi\left[\left(\int_{t=0}^{T} f_t(X_t)\text{d}t\right)^2\right] \leq \frac{4\int_{t=0}^{T}\text{Var}_\pi\left(f_t\right)\text{d}t}{\gamma_{\text{ps}}^{\text{cont}}}. \tag{3.4.27}$$

Now we are going to state Bernstein-type concentration inequalities for reversible and non-reversible processes.

Theorem 3.4.10 (Bernstein inequality for reversible processes)**.** *Let $\{X_t\}_{t\geq 0}$ be a time homogeneous, stationary, reversible Markov process, with distribution $\mathbb{P}$, standard stochastic semigroup $\{\boldsymbol{P}_t\}_{t\geq 0}$, stationary distribution π, and spectral gap γ^{cont}. Let $\{f_t\}_{t\geq 0} \in L^2(\pi)$, satisfying that $|f_t(x) - \mathbb{E}_\pi f_t| \leq C$ for every $t \geq 0$, $x \in \Omega$. Assume that they satisfy the regularity conditions* (3.4.22) *and* (3.4.23)*. Define*

$$S' := \int_{t=0}^{T} f_t(X_t)\text{d}t, \text{ and } V_{S'} := \int_{t=0}^{T}\text{Var}_\pi(f_t)\text{d}t,$$

then

$$\mathbb{P}_\pi(|S' - \mathbb{E}_\pi(S')| \geq r) \leq 2\exp\left(-\frac{r^2\cdot\gamma^{\text{cont}}}{4V_{S'} + 10rC}\right). \tag{3.4.28}$$

Remark. We apply this inequality to M/M/1 queues in Example 3.4.22.

Theorem 3.4.11 (Bernstein inequality for non-reversible Markov processes)**.** *Let*

$\{X_t\}_{t\geq 0}$ be a time homogeneous, stationary Markov process, with stationary distribution π, and pseudo spectral gap $\gamma_{\mathrm{ps}}^{\mathrm{cont}}$. Let $f \in L^2(\pi)$, satisfying that $|f(x) - \mathbb{E}_\pi f| \leq C$ for every $t \geq 0$, $x \in \Omega$. Let

$$S := \int_{t=0}^{T} f(X_t)\mathrm{d}t,$$

then for any $r \geq 0$

$$\mathbb{P}_\pi(|S - \mathbb{E}_\pi(S)| \geq r) \leq 2\exp\left(-\frac{r^2 \cdot \gamma_{\mathrm{ps}}^{\mathrm{cont}}}{8(T + 1/\gamma_{\mathrm{ps}}^{\mathrm{cont}})\mathrm{Var}_\pi(f) + 20rC}\right). \tag{3.4.29}$$

The following theorem generalises this to integrals of the form $\int_{t=0}^{T} f_t(X_t)\mathrm{d}t$.

Theorem 3.4.12. *Let $\{X_t\}_{t\geq 0}$ be a time homogeneous, stationary Markov process, with stationary distribution π, and pseudo spectral gap $\gamma_{\mathrm{ps}}^{\mathrm{cont}}$. Let $\{f_t\}_{t\geq 0} \in L^2(\pi)$, satisfying that $|f_t(x) - \mathbb{E}_\pi f_t| \leq C$ for every $t \geq 0$, $x \in \Omega$. Assume that they satisfy the regularity conditions* (3.4.22) *and* (3.4.23)*. Define*

$$S' := \int_{t=0}^{T} f_t(X_t)\mathrm{d}t, \text{ and } V_{S'} := \int_{t=0}^{T} \mathrm{Var}_\pi(f_t)\mathrm{d}t.$$

Suppose that there is $t_{\mathrm{ps}} \in \mathbb{R}_+$ such that $\gamma_{\mathrm{ps}} = \gamma((\boldsymbol{P}_{t_{\mathrm{ps}}})^*(\boldsymbol{P}_{t_{\mathrm{ps}}}))/t_{\mathrm{ps}}$ (if this does not exist, we can use a limiting argument). Let

$$M := \frac{\int_{t=0}^{t_{\mathrm{ps}}} \left(\sum_{j=0}^{\lfloor (T-t)/t_{\mathrm{ps}} \rfloor} \mathrm{Var}_\pi(f_{t+jt_{\mathrm{ps}}})\right)^{1/2} \mathrm{d}t}{\inf_{t\in[0,t_{\mathrm{ps}}]} \left(\sum_{j=0}^{\lfloor T/t_{\mathrm{ps}} \rfloor} \mathrm{Var}_\pi(f_{t+jt_{\mathrm{ps}}})\right)^{1/2}}.$$

Then for any $r \geq 0$,

$$\mathbb{P}_\pi(|S' - \mathbb{E}_\pi(S')| \geq r) \leq 2\exp\left(-\frac{r^2 \cdot \gamma_{\mathrm{ps}}}{8V_{S'} + 20rC \cdot M/t_{\mathrm{ps}}}\right). \tag{3.4.30}$$

The following theorem is the main results of Lezaud (2001) (which we compare to our results in the following remark).

Theorem 3.4.13 (Theorem 1.1. of Lezaud (2001))**.** *Let P_t be an ergodic Markov semigroup with invariant probability measure π. Assume that its infinitesimal generator L has as simple isolated eigenvalue $\lambda = 0$ and that the initial distribution q has a $L^2(\pi)$ density relatively to the measure π. Let X_s^{sym} be a Markov process with generator $(\boldsymbol{L} + \boldsymbol{L}^*)/2$, $S_T^{\mathrm{sym}} := \int_0^T f(X_s^{\mathrm{sym}})dt$, and*

$$\sigma_{\mathrm{sym}}^2 := \lim_{T\to\infty} T^{-1}\mathrm{Var}_\pi(S_T^{\mathrm{sym}}).$$

Let $f \in D_2(\boldsymbol{L})$ such that $\mathbb{E}_\pi(f) = 0$, and $\|f\|_\infty \le C$. Then for all $r > 0$, $T > 0$,

$$\mathbb{P}_q(T^{-1}S_T \ge r) \le N_q \exp\left\{ -\frac{2Tr^2}{\sigma_{\mathrm{sym}}^2\left(1 + \sqrt{1 + 4Cr/(\gamma_{\mathrm{sym}}\sigma_{\mathrm{sym}}^2)}\right)^2} \right\}, \tag{3.4.31}$$

with $S_T := \int_0^T f(X_s)dt$, γ_{sym} is the spectral gap of $(\boldsymbol{L} + \boldsymbol{L}^*)/2$, and N_q is the $L^2(\pi)$ norm of the density of q related to the stationary distribution π.

Remark 3.4.14. A similar bound is given, by different methods, in Guillin, Léonard, Wu, and Yao (2009). As we can see, for reversible processes, $\sigma_{\mathrm{cont}} = \sigma_{\mathrm{sym}}$, and

$$-\frac{2Tr^2}{\sigma_{\mathrm{cont}}^2\left(1 + \sqrt{1 + 4Cr/(\gamma_{\mathrm{cont}}\sigma_{\mathrm{cont}}^2)}\right)^2} \le -\frac{Tr^2}{2\sigma_{\mathrm{cont}}^2 + 4(C/\gamma_{\mathrm{cont}})r},$$

thus for $r \le \sigma_{\mathrm{cont}}^2\gamma_{\mathrm{cont}}/2C$, this is smaller than $-Tr^2/(4\sigma_{\mathrm{cont}}^2)$, and for $r \ll \sigma_{\mathrm{cont}}^2\gamma_{\mathrm{cont}}/(2a)$, it is essentially $-Tr^2/(2\sigma_{\mathrm{cont}}^2)$. The constant 2 is sharp here because of the asymptotic normality of the empirical average. For reversible processes, this bound is sharper

than Theorem 3.4.10, since it uses the asymptotic variance (in this case, the asymptotic variance σ^2_{cont} equals σ^2_{sym}). However, for non-reversible processes, in general $\sigma^2_{\text{cont}} \neq \sigma^2_{\text{sym}}$, and γ_{sym} is the spectral gap of symmetrized operator $(\boldsymbol{L} + \boldsymbol{L}^*)/2$, which may be 0 even for fast mixing processes. Thus Theorem 3.4.11, involving $\gamma^{\text{cont}}_{\text{ps}}$, can be sharper in this case.

Finally, we present a version of McDiarmid's bounded differences inequality for continuous time Markov processes.

Theorem 3.4.15 (McDiarmid's bounded differences inequality for Markov processes)**.** *Let $\{X_t\}_{0\leq t\leq T}$ be a (not necessarily time homogeneous) Markov process, with X_t taking values in a Polish space Λ_t. Let $\Lambda := \prod_{t=0}^{T} \Lambda_t$. Suppose that its mixing time is given by $\tau^{\text{cont}}(\epsilon)$ (for $0 \leq \epsilon \leq 1$).*

Assume that some function $f : \Lambda \to \mathbb{R}$ satisfies that for some $c : [0,T] \to \mathbb{R}_+$, for every $0 \leq a < b \leq T$, every $x, x' \in \Lambda$ such that $x_t = x'_t$ whenever $t \notin [0,T] \setminus [a,b]$,

$$|f(x) - f(x')| \leq \int_{t=a}^{b} c(t)\mathrm{d}t.$$

Let $\Lambda^{(N)} := \Lambda_0 \times \Lambda_{T/N} \times \ldots \times \Lambda_{T(N-1)/N}$, and define $X^{(N)} \in \Lambda^{(N)}$ as $X^{(N)}_i = X_{(i/N)T}$ for $0 \leq i < N$. Assume that f also satisfies the following regularity conditions: for some sequence of functions $f^{(N)} : \Lambda^{(N)} \to \mathbb{R}$, $N \in \mathbb{N}$,

$$\lim_{N\to\infty} f^{(N)}(X^{(N)}) = f(X) \quad \mathbb{P} \text{ almost surely,} \tag{3.4.32}$$

$$\lim_{N\to\infty} \mathbb{E}\left(f(X^{(N)})\right) = \mathbb{E}\left(f(X)\right). \tag{3.4.33}$$

Let

$$\tau_{\min}^{\text{cont}} := \inf_{0 \le \epsilon < 1} \tau^{\text{cont}}(\epsilon) \cdot \left(\frac{2-\epsilon}{1-\epsilon}\right)^2. \tag{3.4.34}$$

Then for any $r \ge 0$,

$$\mathbb{P}\left(|f(X) - \mathbb{E}f(X)| \ge r\right) \le 2\exp\left(\frac{-2r^2}{\int_{t=0}^{T} c^2(t)\mathrm{d}t \cdot \tau_{\min}^{\text{cont}}}\right). \tag{3.4.35}$$

3.4.3 Extension to non-stationary chains, and unbounded functions

In the previous section, we have stated variance bounds and Bernstein-type inequalities for integrals of the form $\int_{t=0}^{T} f_t(X_t)\mathrm{d}t$, with $\{X_t\}_{0 \le t \le T}$ being a stationary time homogeneous Markov process. 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 Section3.6.3.

The following two propositions are useful to generalise such bounds to non-stationary processes.

Proposition 3.4.16 (Bounds for non-stationary processes)**.** *Let* $X := (X_t)_{0 \le t \le T}$ *be a time homogenous Markov process with state space* Ω*, and stationary distribution* π*. Suppose that* $g(X)$ *is real valued measurable function. Then*

$$\mathbb{P}_q(g(X) \ge r) \le N_q^{1/2} \cdot \left[\mathbb{P}_\pi(g(X) \ge r)\right]^{1/2}, \tag{3.4.36}$$

for any distribution q *on* Ω*. Now suppose that we "burn" observations up to some*

time t_0, and we are interested in bounds on a function h of $(X_t)_{t_0 \le t \le T}$. Firstly,

$$\mathbb{P}_q(h((X_t)_{t_0 \le t \le T}) \ge r) \le N_{q\boldsymbol{P}_{t_0}}^{1/2} \cdot \left[\mathbb{P}_\pi(h((X_t)_{t_0 \le t \le T}) \ge r)\right]^{1/2}, \tag{3.4.37}$$

moreover,

$$\mathbb{P}_q(h((X_t)_{t_0 \le t \le T}) \ge r) \le \mathbb{P}_\pi(h((X_t)_{t_0 \le t \le T}) \ge r) + d_{\mathrm{TV}}\left(q\boldsymbol{P}_{t_0}, \pi\right). \tag{3.4.38}$$

Proposition 3.4.17 (Further bounds for non-stationary processes)**.** *In Proposition 3.4.16, $N_{q\boldsymbol{P}_{t_0}}$ can be further bounded. For reversible processes, we have*

$$N_{q\boldsymbol{P}_{t_0}} \le 1 + (N_q - 1) \cdot (1 - \gamma^{\mathrm{cont}})^{2t_0}, \tag{3.4.39}$$

while for non-reversible processes,

$$N_{q\boldsymbol{P}_{t_0}} \le 1 + (N_q - 1) \cdot (1 - \gamma_{\mathrm{ps}}^{\mathrm{cont}})^{2(t_0 - 1/\gamma_{\mathrm{ps}}^{\mathrm{cont}})}. \tag{3.4.40}$$

Similarly, $d_{\mathrm{TV}}\left(q\boldsymbol{P}_{t_0}, \pi\right)$ can be further bounded too. For reversible processes, we have, by (3.4.11)*,*

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}_{t_0}, \pi\right) \le \frac{1}{2}(1 - \gamma^{\mathrm{cont}})^{t_0} \cdot \sqrt{N_q - 1},$$

For non-reversible processes, by (3.4.14)*,*

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}_{t_0}, \pi\right) \le \frac{1}{2}(1 - \gamma_{\mathrm{ps}}^{\mathrm{cont}})^{(t_0 - 1/\gamma_{\mathrm{ps}}^{\mathrm{cont}})/2} \cdot \sqrt{N_q - 1},$$

Finally, for uniformly ergodic Markov processes,

$$d_{\mathrm{TV}}\left(q\boldsymbol{P}_{t_0}, \pi\right) \leq \inf_{0\leq\epsilon<1} \epsilon^{\lfloor t_0/\tau^{\mathrm{cont}}(\epsilon)\rfloor} \leq 2^{-\lfloor t_0/t_{\mathrm{mix}}^{\mathrm{cont}}\rfloor}. \tag{3.4.41}$$

The Bernstein-type inequalities will assume boundedness of the functions that we integrate. In order to generalise such bounds to unbounded functions, we can use truncation. For $a, b \in \mathbb{R}$, $a < b$, define $\mathcal{T}_{[a,b]}(x) = x \cdot \mathbb{1}[x \in [a, b]] + a \cdot \mathbb{1}[x < a] + b \cdot \mathbb{1}[x > b]$ (as in Section 3.3.1), then we have following proposition (continuous analogue of Proposition 3.3.14).

Proposition 3.4.18 (Truncation for Markov processes)**.** *Let $(X_t)_{0\leq t\leq T}$ be a time homogeneous, stationary Markov process with stationary distribution π, and Polish state space Ω. Let $f : \Omega \to \mathbb{R}$ be a measurable function. Then for any $a < b$, for any r,*

$$\begin{aligned}\mathbb{P}_\pi\left(\int_{t=0}^{T} f(X_t)\mathrm{d}t \geq r\right) &\leq \mathbb{P}_\pi\left(\int_{t=0}^{T} \mathcal{T}_{[a,b]}(f(X_t))\mathrm{d}t \geq r\right) \\ &+ \mathbb{P}_\pi\left(\inf_{0\leq t\leq T} f(X_t) < a\right) + \mathbb{P}_\pi\left(\sup_{0\leq t\leq T} f(X_t) > b\right).\end{aligned}$$

Remark 3.4.19. A similar bound can be given for integrals of the form $\int_{t=0}^{T} f_t(X_t)\mathrm{d}t$. Example 3.4.22 shows an application to M/M/1 queues.

3.4.4 Applications

Example 3.4.20 (Convergence of empirical distribution for Markov processes)**.** Let $\{X_t\}_{0\leq t\leq T}$ be a uniformly ergodic Markov chain with countable state space Λ, unique stationary distribution π, and mixing time $t_{\mathrm{mix}}^{\mathrm{cont}}$. We denote its empirical distribution

by $\pi_{em}(x) := \frac{1}{T}\int_{t=0}^{T} \mathbb{1}[X_i = x]$. In Examples 3.2.17 and 3.2.17, we have analysed the the total variational distance of the empirical distribution and the stationery distribution, and shown concentration inequalities for it. The following proposition proves an analogous result for Markov processes.

Proposition 3.4.21. *Denote*

$$d(\{X_t\}_{0\le t\le T}) := d_{\mathrm{TV}}(\pi_{em}(x), \pi),$$

then for any $r \ge 0$*,*

$$\mathbb{P}(|d(\{X_t\}_{0\le t\le T}) - \mathbb{E}(d)| \ge r) \le 2\exp\left(-\frac{t^2 \cdot T}{4.5 t_{\mathrm{mix}}^{\mathrm{cont}}}\right).$$

For stationary, reversible processes,

$$\mathbb{E}(d) \le \sum_{x\in\Lambda} \min\left(\sqrt{\frac{2\pi(x)}{T\gamma^{\mathrm{cont}}}}, \pi(x)\right). \tag{3.4.42}$$

In the case of stationary, non-reversible processes, (3.4.42) *holds with* γ^{cont} *replaced by* $\gamma_{\mathrm{ps}}^{\mathrm{cont}}/2$*.*

Proof. The proof is similar to the proofs in Examples 3.2.17 and 3.2.17, except that we use Theorems 3.4.15, 3.4.7 and 3.4.9. □

Example 3.4.22 (Empirical averages for M/M/1 queues)**.** Then M/M/1 queue is a simple single server queue model. Customers arrive with exponential interarrival times with mean $1/\lambda$, and they are served in order of arrival, by a single server with exponential service times with mean $1/\mu$. Denote the number of customers in the queue at time t by X_t. Then X_t is a continuous time Markov process, with state

space $\Omega = \mathbb{N}$. If $\rho := \lambda/\mu < 1$, then it is reversible with respect to the stationary distribution $\pi(x) = (1-\rho)\rho^x$, $x \in \mathbb{N}$ (geometric distribution, with parameter $1-\rho$). The mean and variance of this distribution are given by

$$M := \frac{\rho}{1-\rho}, \text{ and } V := \rho/(1-\rho)^2. \tag{3.4.43}$$

In the following, we will suppose that our process is stationary, that is $\rho < 1$, and $X_0 \sim \pi$. The M/M/1 queue has been well studied, in particular, it is proven in Karlin and McGregor (1958) that if $\rho < 1$, then

$$\gamma^{\text{cont}} = (\mu^{1/2} - \lambda^{1/2})^2. \tag{3.4.44}$$

Consider the average number of customers in the queue up to time T,

$$A_T := \frac{1}{T}\int_{t=0}^{T} X_t \mathrm{d}t. \tag{3.4.45}$$

This is the continuous time empirical average of the unbounded function $f(x) = x$. By Theorem 3.4.7, we can bound the variance of this quantity as

$$\text{Var}(A_T) \leq \frac{2V\gamma^{\text{cont}}}{T}. \tag{3.4.46}$$

The following proposition states a concentration inequality for A_T.

Proposition 3.4.23. *For any $s \geq 0$, let*

$$B(s) := \max\left((V/5s)\cdot\left(-1+\sqrt{1+\frac{5}{2V^2}\cdot s^3 T\cdot\frac{\gamma^{\text{cont}}}{\log(1/\rho)}}\right), \frac{4\log(\lambda T)}{\log(1/\rho)}\right).$$

Then

$$\mathbb{P}\left[|A_T - M| \geq \frac{\rho^{B(s)+1}}{1-\rho} + s\right] \leq 4(\lambda T)\log(1/\rho)B(s)\exp\left[-\frac{s^2 T\gamma^{\text{cont}}}{4V + 10B(s)\cdot s}\right]. \tag{3.4.47}$$

Remark 3.4.24. If s is small ($s \leq (\log(1/\rho)V^2/(\gamma^{\text{cont}}T))^{1/3}$), then the exponential term is of the form $\exp(-s^2T\gamma^{\text{cont}}/(6V))$, while for larger s, it is of the form $\exp(-\mathcal{O}(\sqrt{sT}))$.

We will prove this proposition using a truncation argument. The following lemma will be used in the proof.

Lemma 3.4.25. *Suppose that $(X_t)_{0\leq t\leq T}$ is a stationary M/M/1 queue. Let $Y_T := \sup_{0\leq t\leq T} X_t$. Then for $T \geq 2/\lambda$, $b \geq 4\log(\lambda T)/\log(1/\rho)$,*

$$\mathbb{P}(Y_T > b) \leq 2(\lambda T)\rho^{b/2}b\log(1/\rho).$$

Proof. For $N \geq 1$ positive integer, let $Y_T^{(N)} := \max_{0\leq i<N} X_{\frac{i}{N}T}$ be the "discretised approximation" of Y_T. Then obviously, we have $Y_T^{(N)} \leq Y_T$. Notice that the supremum in the definition of Y_T is achieved at one of the arrival times, denote the smallest such by time by t_s. Then between t_s and the previous arrival time there can be no one served. Therefore if we choose N such that T/N is smaller than the shortest interarrival time up to time T, then $Y_T \leq Y_T^{(N)} + 1$. Denote the event that the shortest interarrival time up to time T is shorter than T/N by $E_{T/N}$.

Now we need to get an upper bound on $\mathbb{P}(E_{T/N})$. The total number of arrivals to time T can be shown to be Poisson distributed with parameter λT. Denote this distribution by $\mu_{\lambda T}^{\text{Poi}}$. Since the exponential distribution of parameter λ has density

$f_\lambda(x) = \lambda \exp(-\lambda x) \leq \lambda$ for $x \geq 0$, the probability of an interarrival time being shorter than T/N is smaller than $\lambda T/N$. Therefore, the probability that amongst the first M interarrival times, there is at least one shorter than T/N is smaller then $M\lambda T/N$. This means that

$$\mathbb{P}(E_{T/N}) \leq \mu_{\lambda T}^{\text{Poi}}[(M, \infty)] + \frac{M\lambda T}{N}.$$

for any $M \in \mathbb{N}$. The moment generating function of the Poisson distribution $\mu_{\lambda T}^{\text{Poi}}$ function is known to be $\exp(\lambda T(e^\theta - 1))$. When $\theta = \log(1 + 1/(\lambda T))$, this equals e, so by Markov's inequality, we obtain

$$\mu_{\lambda T}^{\text{Poi}}[(M, \infty)] \leq e \cdot \exp[-M \log(1 + 1/(\lambda T))].$$

Using this in our bound on $\mathbb{P}(E_{T/N})$, we obtain

$$\mathbb{P}(E_{T/N}) \leq e \cdot \exp[-M \log(1 + 1/(\lambda T))] + \frac{M\lambda T}{N}.$$

By setting $M := 2\lambda T \log(eN/(2(\lambda T)^2))$, we obtain that

$$\mathbb{P}(E_{T/N}) \leq 4\frac{(\lambda T)^2}{N} \log\left(\frac{N}{(\lambda T)^2}\right) \text{ for } N \geq e(\lambda T)^2.$$

Now $Y_T^{(N)} = \max_{0 \leq i < N} X_{\frac{i}{N}T}$, and $X_{\frac{i}{N}T}$ is distributed geometrically with parameter $1 - \rho$, so it is easy to see that for any $b \in \mathbb{N}$,

$$\mathbb{P}(Y_T^{(N)} > b) \leq N \sum_{i=b+1}^{\infty} \rho^i (1 - \rho) = N\rho^{b+1}.$$

Moreover, we know that outside of the event $E_{N/T}$, $Y_T \leq Y_T^{(N)} + 1$, so for $N \geq e(\lambda T)^2$,

$$\mathbb{P}(Y_T > b) \leq N\rho^b + 4\frac{(\lambda T)^2}{N} \log\left(\frac{N}{(\lambda T)^2}\right).$$

Now the statement of the lemma follows by setting $N = 2\lambda T/(\rho^{b/2})$. □

Now we are ready to prove our concentration bound.

Proof of Proposition 3.4.23. In order to get a concentration inequality for A_T, we are going to apply the truncation argument of Proposition 3.4.18. For any $b \in \mathbb{N}$, $s \geq 0$,

$$\begin{aligned} &\mathbb{P}_\pi\left(\left|\frac{1}{T}\int_{t=0}^{T} X_t \mathrm{d}t - M\right| \geq M - \mathbb{E}_\pi(\mathcal{T}_{[0,b]}(X_0)) + s\right) \leq \\ &\mathbb{P}_\pi\left(\left|\frac{1}{T}\int_{t=0}^{T} \mathcal{T}_{[0,b]}(X_t)\mathrm{d}t - \mathbb{E}_\pi(\mathcal{T}_{[0,b]}(X_0))\right| \geq s\right) + \mathbb{P}_\pi\left(\sup_{0\leq t\leq T} X_t > b\right). \end{aligned} \tag{3.4.48}$$

A simple calculation shows that $\mathbb{E}_\pi(\mathcal{T}_{[0,b]}(X_t)) = \frac{\rho}{1-\rho} - \frac{\rho^{b+1}}{1-\rho}$ and $\mathrm{Var}_\pi \mathcal{T}_{[0,b]}(X_t) \leq \mathrm{Var}_\pi X_t = V$. By Bernstein's inequality for reversible processes (Theorem 3.4.10), we have

$$\mathbb{P}_\pi\left(\left|\frac{1}{T}\int_{t=0}^{T} \mathcal{T}_{[0,b]}(X_t)\mathrm{d}t - \mathbb{E}_\pi(\mathcal{T}_{[0,b]}(X_0))\right| \geq s\right) \leq 2\exp\left(-\frac{s^2\gamma^{\mathrm{cont}}T}{4V + 10sb}\right).$$

By Lemma 3.4.25, it follows that for $T \geq 2/\lambda$, $b \geq 4\log(\lambda T)/\log(1/\rho)$,

$$\mathbb{P}_\pi\left(\sup_{0\leq t\leq T} X_t > b\right) \leq 2(\lambda T) b \log(1/\rho)\rho^{b/2}.$$

The statement of the proposition follows from (3.4.48), by setting $b = B(s)$, and noticing that $\rho^{B(s)/2} \leq \exp\left(-\frac{s^2\gamma^{\mathrm{cont}}T}{4V+10sB(s)}\right)$. □

A similar argument is possible for queues involving k servers ($M/M/k$ queues). The spectral gaps for such queues are computed in Karlin and McGregor (1958). Note also that in the case of an infinity number of servers (the so called $M/M/\infty$ queue), Joulin and Ollivier (2010) proves an exponential concentration bound for empirical averages of Lipschiz functions (such as the average number of persons in the queue up to time T) using the coarse Ricci curvature approach (see also Joulin (2009), and Guillin, Léonard, Wu, and Yao (2009)).

3.5 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 (3.2.4) is Proposition 1 of Marton (1996b), 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 (3.2.2)). The proof is based on the transportation cost inequality method. Marton (1996a, 1997, 1998) 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. Samson (2000) 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, Marton (2003) further develops the results of Samson (2000), and introduces the coupling structure that we call Marton

coupling in this chapter. There are further extensions of this method to more general distances, and mixing conditions, see Rio (2000), Djellout, Guillin, and Wu (2004), and Wintenberger (2012). Alternative, simpler approaches to show McDiarmid-type inequalities for dependent random variables were developed in Chazottes, Collet, Külske, and Redig (2007) (using an elementary martingale-type argument) and Kontorovich and Ramanan (2008) (using martingales and linear algebraic inequalities). For time homogeneous Markov chains, their results are similar to Proposition 1 of Marton (1996b).

In this chapter, 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 Gillman (1998), which shows a Hoeffding-type inequality for reversible chains. The method was further developed in Lezaud (1998a), where Bernstein-type inequalities are obtained. A sharp version of Hoeffding's inequality for reversible chains was proven in León and Perron (2004).

The second popular approach in the literature is by regeneration-type minorisation conditions, see Glynn and Ormoneit (2002) and Douc, Moulines, Olsson, and van Handel (2011) for Hoeffding-type inequalities, and Adamczak and Bednorz (2012) 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, making them harder to apply in practice than spectral methods.

In this chapter, we have sharpened the bounds of Lezaud (1998a). 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 Lezaud (1998a).

3.6 Proofs

3.6.1 Proofs by Marton couplings

Proof of Proposition 3.2.4. 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 $1 \le i \le n$, 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 3.2.6. Now using the Markov property, it is easy to show that for any $1 \le i < j \le n$, the total variational distance of $Ł(\hat{X}_j, \dots \hat{X}_n | \hat{X}_1 = \hat{x}_1, \dots, \hat{X}_i = \hat{x}_i)$ and $Ł(\hat{X}_j, \dots \hat{X}_n | \hat{X}_1 = \hat{x}_1, \dots, \hat{X}_{i-1} = \hat{x}_{i-1}, \hat{X}_i = \hat{x}_i')$ equals to the total variational

distance of $Ł(X_j|\hat{X}_1 = \hat{x}_1, \ldots, \hat{X}_i = \hat{x}_i)$ and $Ł(X_j|\hat{X}_1 = \hat{x}_1, \ldots, \hat{X}_{i-1} = \hat{x}_{i-1}, \hat{X}_i = \hat{x}_i')$, 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 3.2.9 (due to Devroye and Lugosi (2001)).

Lemma 3.6.1. *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

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

Proof of Theorem 3.2.9. We prove this result based on the martingale approach of Chazottes, Collet, Külske, and Redig (2007) (a similar proof is possible using the method of Kontorovich (2007)). Let $\hat{f}(\hat{X}) := f(X)$, then it satisfies that for every $\hat{x}, \hat{y} \in \hat{\Lambda}$,

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

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

$$\log \mathbb{E}\left(e^{\lambda(f(X) - \mathbb{E}f(X))}\right) \leq \frac{\lambda^2 \cdot \|\Gamma \cdot c\|^2}{8} \tag{3.6.1}$$

under the assumption that there is a Marton coupling for X with mixing matrix Γ. By applying this inequality to $\hat{X}$, (3.2.5) follows.

Now we will show (3.6.1). Let us define $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$ for $i \leq N$, and write $f(X) - \mathbb{E}f(X) = \sum_{i=1}^N V_i(X)$, with

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

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, Chazottes, Collet, Külske, and Redig (2007) defines a coupling between the distributions

$$\mathcal{L}(X_{i+1}, \ldots, X_N | X_1, \ldots, X_{i-1}, X_i = a),$$
$$\mathcal{L}(X_{i+1}, \ldots, X_N | X_1, \ldots, X_{i-1}, X_i = b)$$

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

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

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

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

Now using Lemma 3.6.1 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

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

By taking the product of these, we obtain (3.6.1), and as a consequence, (3.2.5). The tail bounds follow by Markov's inequality. □

Proof of Corollary 3.2.11. We use the Marton coupling of Proposition 3.2.4. By the simple fact that $\|\Gamma\| \le \sqrt{\|\Gamma\|_1\|\Gamma\|_\infty}$, we have $\|\Gamma\| \le 2/(1-\epsilon)$, so applying Theorem 3.2.9 and taking infimum in ϵ proves the result. □

3.6.2 Proofs by spectral methods

Proof of Proposition 3.3.2. The proof of the first part is similar to the proof of Proposition 30 of Ollivier (2009). 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

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

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 $0 \le \epsilon < 1$, it is easy to see that $\|\boldsymbol{M}_\infty^{\tau(\epsilon)}\|_\infty \le 2\epsilon$, and for $l \ge 1$, $\tau(\epsilon^l) \le l\tau(\epsilon)$, thus $\|\boldsymbol{M}_\infty^{l\tau(\epsilon)}\|_\infty \le 2\epsilon^l$. Therefore, we can show that

$$\lim_{k\to\infty} \|\boldsymbol{M}_\infty^k\|_\infty^{1/k} \le \epsilon^{1/\tau(\epsilon)}. \tag{3.6.2}$$

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

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

Now we turn to the proof of (3.3.6). For Markov chains on finite state spaces, (3.3.6) is a reformulation of Theorem 2.7 of Fill (1991) (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 3.3.3. 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 $0 \le \epsilon < 1$. 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 (3.3.10). Note that for any $k \geq 1$,

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

Now using Theorem 2.7 of Fill (1991) with $\boldsymbol{M} = (\boldsymbol{P}^*)^k\boldsymbol{P}^k$, we obtain

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

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

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

Proof of Theorem 3.3.4. 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

$$\mathbb{E}_\pi\left(f(X_i)\sum_{j=1}^n f(X_j)\right) = \left\langle f, \left(\sum_{j=1}^n (\boldsymbol{P}-\boldsymbol{\pi})^{|j-i|}\right) f \right\rangle_\pi, \tag{3.6.3}$$

where

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

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 $1 \leq i \leq n$,

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

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

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

Summing up in i leads to (3.3.14).

Now we turn to the proof of (3.3.15). Summing up (3.6.3) in i leads to

$$\mathbb{E}\left(\left(\sum_{i=1}^{n}f(X_i)\right)^2\right) = \left\langle f, \left[(2n\boldsymbol{I} - 2\left(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi})^{n-1}\right)(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi}))^{-1}\right)\right. \quad (3.6.4)$$
$$\left.\cdot\,(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi}))^{-1} - n\boldsymbol{I}\right] f\Big\rangle_{\pi},$$

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

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

Now we turn to the proof of (3.3.16). For stationary chains, for $1 \leq i, j \leq n$,

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

and thus for any $1 \leq i, j \leq n$, $\mathbb{E}(f_i(X_i)f_j(X_j)) \leq \frac{1}{2}\mathbb{E}_\pi(f_i^2 + f_j^2)(1-\gamma^*)^{|i-j|}$. Summing up in i and j proves (3.3.16). □

Proof of Theorem 3.3.6. Without loss of generality, we assume that $\mathbb{E}_\pi(f) = 0$, and $\mathbb{E}_\pi(f_i) = 0$ for $1 \leq i \leq n$. Now for $1 \leq i, j \leq n$,

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

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

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

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 $\mathbb{E}(f(X_i)f(X_j)) \leq V_f(1 - k\gamma_{\text{ps}})^{\frac{1}{2}\lceil \frac{j-i}{k_{\text{ps}}} \rceil}$. By summing up

in i and j, and noticing that

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

we can deduce (3.3.17). By the definition of $\sigma^2_{\rm as}$, it follows that

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

and by comparing this with (3.6.4), we have

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

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

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

Optimizing in k gives $\|(\boldsymbol{I}-(\boldsymbol{P}-\boldsymbol{\pi}))^{-1}\|_{2,\pi} \leq 2/\gamma_{\rm ps}$, and (3.3.18) follows. Finally, the proof of (3.3.19) 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 3.6.2. *Let $X_1, \ldots, X_n$ 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*,*

$$\mathbb{E}_\pi(\exp(\theta S)) = \left\langle \mathbf{1}, (e^{\theta \boldsymbol{D}_f} \boldsymbol{P})^n \mathbf{1} \right\rangle_\pi \leq \|e^{\theta \boldsymbol{D}_f/2} \boldsymbol{P} e^{\theta \boldsymbol{D}_f/2}\|_{2,\pi}^{n-1} \|e^{\theta f/2}\|_{2,\pi}^2, \tag{3.6.5}$$

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

$$\begin{aligned} \mathbb{E}_\pi(\exp(\theta S')) &= \left\langle \mathbf{1}, \left(e^{\theta \boldsymbol{D}_{f_1}} \boldsymbol{P}\right) \cdot \ldots \cdot \left(e^{\theta \boldsymbol{D}_{f_n}} \boldsymbol{P}\right) \mathbf{1} \right\rangle_\pi \\ &= \left\langle \mathbf{1}, \left(\boldsymbol{P} e^{\theta \boldsymbol{D}_{f_1}}\right) \cdot \ldots \cdot \left(\boldsymbol{P} e^{\theta \boldsymbol{D}_{f_n}}\right) \mathbf{1} \right\rangle_\pi \\ &\leq \|\boldsymbol{P} e^{\theta \boldsymbol{D}_{f_1}}\|_{2,\pi} \cdot \ldots \cdot \|\boldsymbol{P} e^{\theta \boldsymbol{D}_{f_n}}\|_{2,\pi}. \end{aligned} \tag{3.6.6}$$

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

Lemma 3.6.3. *Suppose that* $f \in L^2(\pi)$*,* $-1 \leq f \leq 1$*,* $\mathbb{E}_\pi(f) = 0$*, then for reversible* $\boldsymbol{P}$*, for* $0 < \theta < \gamma/10$*, we have*

$$\|e^{\theta \boldsymbol{D}_f} \boldsymbol{P} e^{\theta \boldsymbol{D}_f}\|_{2,\pi} \leq 1 + \frac{4V_f}{\gamma} \cdot \theta^2 \cdot \left(1 - \frac{10\theta}{\gamma}\right)^{-1}, \text{ and} \tag{3.6.7}$$

$$\|e^{\theta \boldsymbol{D}_f} \boldsymbol{P} e^{\theta \boldsymbol{D}_f}\|_{2,\pi} \leq 1 + 2(\sigma_{\text{as}}^2 + 0.8V_f) \cdot \theta^2 \cdot \left(1 - \frac{10\theta}{\gamma}\right)^{-1}, \tag{3.6.8}$$

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

Proof. (3.6.7) is proven in Lezaud (1998b) (pages 47 and 97), see also Lezaud (1998a). We prove (3.6.8) 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 Lezaud (1998b)). 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 Lezaud (1998b), using the theory of linear perturbations, for $0 \leq r \leq \gamma/3$, we have

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

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

$$\beta^{(n)} = \sum_{p=1}^{n}\frac{1}{p}\sum_{\substack{\nu_1+\ldots+\nu_p=n \\ k_1+\ldots+k_p=p-1 \\ v_i\geq 1, k_j\geq 0}}\frac{1}{\nu_1!\cdots\nu_p!}\left\langle f^{\nu_1'}, \boldsymbol{Z}^{(k_1')}\boldsymbol{P}\boldsymbol{D}_f^{\nu_2}\cdots\boldsymbol{Z}^{(k_{p-1}')}\boldsymbol{P}f^{\nu_p'}\right\rangle_\pi. \qquad (3.6.9)$$

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 Lezaud (1998b), $\langle f, \boldsymbol{Z} f\rangle_\pi = \sigma_{\text{as}}^2 + (1/2)\langle f, f\rangle_\pi$, thus $\beta^{(2)} = \sigma_{\text{as}}^2$. 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 $\langle f, \boldsymbol{Z}\boldsymbol{P} f^2\rangle_\pi \leq \frac{V_f}{\gamma}$, $\frac{1}{6}\langle f, \boldsymbol{D}_f f\rangle_\pi \leq \frac{1}{6} V_f$,

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

thus $|\beta^{(3)}| \leq \sigma_{\text{as}}^2/\gamma + (3/2)V_f/\gamma + (1/6)V_f$. Suppose now that $n \geq 4$. First, if $p = n$, then $\nu_1 = \ldots = \nu_p = 1$, thus each such term in (3.6.9) 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

$$\begin{aligned}&\left|\left\langle Z^{1/2} f, \boldsymbol{Z}^{k_1'-1/2}\boldsymbol{P}\boldsymbol{D}_f \cdots \boldsymbol{Z}^{(k_{n-1}')}\boldsymbol{P} D_f \boldsymbol{P}\boldsymbol{Z}^{k_{n-1}'-1/2}(\boldsymbol{Z}^{1/2} f)\right\rangle_\pi\right| \\ &\leq \frac{\langle f, \boldsymbol{Z} f\rangle_\pi}{2\gamma^{n-2}} \leq \frac{\sigma_{\text{as}}^2 + V_f}{2\gamma^{n-2}}.\end{aligned}$$

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

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

Now by (1.11) on page 20 of Lezaud (1998b), 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 Lezaud (1998b), for $n \ge 3$, $D(n) \le 5^{n-2}$. Thus for $n \ge 4$, we have

$$\begin{aligned} |\beta^{(n)}| &\le \frac{4^{n-1}}{n\sqrt{(n-1)\pi}}\frac{\sigma_{\text{as}}^2 + V_f}{2\gamma^{n-2}} + \frac{5^{n-3}}{2\gamma^{n-2}} V_f \\ &\le \frac{5^{n-2}}{\gamma^{n-2}}\left(\frac{\sigma_{\text{as}}^2 + V_f}{2} \cdot \frac{1}{4} + \frac{V_f}{10}\right) \le \frac{5^{n-2}}{\gamma^{n-2}}\left(\frac{\sigma_{\text{as}}^2 + 0.8V_f}{2}\right). \end{aligned} \tag{3.6.10}$$

By comparing this with our previous bounds on $\beta^{(2)}$ and $\beta^{(3)}$, we can see that (3.6.10) holds for every $n \ge 2$. By summing up, we obtain

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

and substituting $r = 2\theta$ gives (3.6.8). □

Proof of Theorem 3.3.7. We can assume, without loss of generality, that $C = 1$. First,

we will prove the bounds for S, then for S'.

By (3.6.5), we have

$$\mathbb{E}_\pi(\exp(\theta S)) \leq \|e^{\theta \boldsymbol{D}_f/2}\boldsymbol{P}e^{\theta \boldsymbol{D}_f/2}\|_2^{n-1} \cdot \mathbb{E}_\pi\left(e^{\theta f}\right). \tag{3.6.11}$$

By (3.6.7), and (3.6.8), we have that for $0 \leq \theta \leq \gamma/5$,

$$\|e^{\theta \boldsymbol{D}_f/2}\boldsymbol{P}e^{\theta \boldsymbol{D}_f/2}\|_{2,\pi} \leq \exp\left(\frac{V_f}{\gamma} \cdot \theta^2 \cdot \left(1 - \frac{5\theta}{\gamma}\right)^{-1}\right), \text{ and} \tag{3.6.12}$$

$$\|e^{\theta \boldsymbol{D}_f/2}\boldsymbol{P}e^{\theta \boldsymbol{D}_f/2}\|_{2,\pi} \leq \exp\left(\frac{\sigma_{\text{as}}^2 + 0.8V_f}{2} \cdot \theta^2 \cdot \left(1 - \frac{5\theta}{\gamma}\right)^{-1}\right). \tag{3.6.13}$$

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

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

and it is also easy to show that this can be indeed further bounded by the right hand sides of (3.6.12) and (3.6.13). Therefore, we obtain that for $0 \leq \theta \leq \gamma/5$,

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

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

Now the bounds (3.3.21) and (3.3.20) 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 $K = 0.5\sigma_{\text{as}}^2 + 0.4V_f$.

Now we are going to prove (3.3.22). Firstly, by (3.6.6), we have

$$\mathbb{E}_\pi(\exp(\theta S')) \le \|\boldsymbol{P}e^{\theta \boldsymbol{D}_{f_1}}\|_{2,\pi} \cdot \ldots \cdot \|\boldsymbol{P}e^{\theta \boldsymbol{D}_{f_n}}\|_{2,\pi}. \tag{3.6.14}$$

Now for $0 \le \theta \le \gamma(\boldsymbol{P}^2)/10$, each of these terms can be further bounded by (3.6.7) as

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

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

$$\mathbb{E}_\pi(\exp(\theta S')) \le \exp\left(\frac{2V_{S'}}{\gamma(\boldsymbol{P}^2)} \cdot \theta^2 \cdot \left(1 - \frac{10\theta}{\gamma(\boldsymbol{P}^2)}\right)^{-1}\right), \tag{3.6.15}$$

and (3.3.22) follows by Markov's inequality. □

Proof of Theorem 3.3.9. We will treat the general case concerning S' first. The proof is based on a trick of Janson (2004). First, we divide the sequence $f_1(X_1), \ldots, f_n(X_n)$ into k_{ps} parts,

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

Denote the sums of each part by $S'_1, \ldots, S'_{k_{\text{ps}}}$, then $S' = \sum_{i=1}^{k_{\text{ps}}} S'_i$. By Yensen's inequality, for any weights $0 \le p_1, \ldots, p_{k_{\text{ps}}} \le 1$ with $\sum_{i=1}^{k_{\text{ps}}} p_i = 1$,

$$\mathbb{E}_\pi \exp(\theta S') \le \sum_{i=1}^{k_{\text{ps}}} p_i \mathbb{E}_\pi \exp((\theta/p_i) \cdot S'_i). \tag{3.6.16}$$

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 (3.6.6) 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 (3.6.7), 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

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

These bounds hold for every $1 \le i \le k_{\rm ps}$. Setting p_i in (3.6.16) as

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

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

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

and (3.4.30) follows by Markov's inequality. In the case of (3.3.23), we have

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

which implies that

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

Now (3.3.23) follows by Markov's inequality and $k_{\rm ps}\lceil n/k_{\rm ps}\rceil \leq n + 1/\gamma_{\rm ps}$. □

Proof of Proposition 3.3.12. Inequalities (3.3.25) and (3.3.26) follow by writing

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

and then applying Cauchy-Schwartz inequality. Inequality (3.3.27) follows by noticing that by the Markov property, the two distributions

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

have total variational distance equal to the total variational distance of

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

Proof of Proposition 3.3.13. Inequalities (3.3.28) and (3.3.29) follow from (2.11) on page 68 of Fill (1991), similarly to the proof of Proposition 3.3.3 (by noticing that the χ^2 distance can be written as $N_q - 1$). Finally, (3.4.41) follows from the definition of $\tau(\epsilon)$ and t_{mix}. $\square$

Proof of Proposition 3.3.14. This follows by a straightforward coupling argument. The details are left to the reader. $\square$

3.6.3 Proofs for continuous time Markov processes

Proofs of Propositions 3.4.5 and 3.4.6. Notice that $\boldsymbol{P}_{\tau(\epsilon)} = e^{\tau(\epsilon)\boldsymbol{L}}$ (as operators on $\mathcal{D}_2(\boldsymbol{L})$), and thus $\gamma(\boldsymbol{P}_{\tau(\epsilon)}) = e^{\tau(\epsilon)\gamma^{\text{cont}}}$. From here, the proof is similar to the proof of Propositions 3.3.2 and 3.3.3, and it is left to the reader as an exercise. Note that in the reversible case, one needs to use Theorem 2.14 of Fill (1991) for proving (3.4.11) (in the non-reversible case, we can still use Theorem 2.7). $\square$

Proofs of Propositions 3.4.16 and 3.4.17. The proofs are similar to the proofs of Propositions 3.3.12 and 3.3.13 $\square$

Proof of Proposition 3.4.18. This follows by a straightforward coupling argument. The details are left to the reader. $\square$

Proof of Theorem 3.4.8. This result follows by applying (3.3.16) of Theorem 3.3.4 to

bound the variance of sums of the form

$$\frac{T-t_0}{N}\sum_{k=1}^{N} f_{t_0+\frac{T-t_0}{N}\cdot k}\left(X_{t_0+\frac{T-t_0}{N}\cdot k}\right),$$

and then taking the limit as $N \to \infty$. The details are left to the reader. Notice that in continuous time, there is no difference between the spectral gap and absolute spectral gap (since all the eigenvalues of the generator $\boldsymbol{L}$ are negative). □

Proof of Theorem 3.4.7. This is similar to the proof of Theorem 3.4.8. □

Proof of Theorem 3.4.9. The proof is similar to the previous one, except here we use Theorem 3.3.6. □

Proof of Theorem 3.4.10. By (3.4.22), it is sufficient to consider tail bounds of the form

$$\mathbb{P}_\pi\left(\frac{T}{N}\sum_{k=1}^{N} f_{\frac{T}{N}\cdot k}\left(X_{\frac{T}{N}\cdot k}\right) \geq r\right).$$

Now for $\{X_{\frac{T}{N}\cdot k}\}_{k\geq 1}$ is a reversible Markov chain with transition kernel $P_{T/N}$, so we can apply Theorem 3.3.7. Using the fact that $\boldsymbol{P}_{T/N} = e^{T/N\cdot\boldsymbol{L}}$, it follows that $\gamma(\boldsymbol{P}_{T/N}) = \gamma^*(\boldsymbol{P}_{T/N})$, and thus by (3.3.21),

$$\begin{aligned}&\mathbb{P}_\pi\left(\frac{T}{N}\sum_{k=1}^{N} f_{\frac{T}{N}\cdot k}\left(X_{\frac{T}{N}\cdot k}\right) \geq r\right)\\ &\leq 2\exp\left(-\frac{r^2\cdot(2\gamma(\boldsymbol{P}_{T/N}) - (\gamma(\boldsymbol{P}_{T/N}))^2)\cdot N^2/T^2}{8\sum_{k=1}^{N}\operatorname{Var}\left(f_{\frac{T}{N}\cdot k}\left(X_{\frac{T}{N}\cdot k}\right)\right) + 20rC\cdot N/T}\right).\end{aligned}$$

Now we can see that

$$\gamma(\boldsymbol{P}_{t/N}) = \gamma(e^{T/N\cdot\boldsymbol{L}}) = \frac{T}{N}\gamma^{\text{cont}} + o(1/N), \tag{3.6.17}$$

thus, using (3.4.23), we obtain that

$$\lim_{N\to\infty} \frac{\gamma(\boldsymbol{P}_{T/N}) - (\gamma(\boldsymbol{P}_{T/N}))^2) \cdot N^2/T^2}{8\sum_{k=1}^{N} \operatorname{Var}\left(f_{\frac{T}{N}\cdot k}\left(X_{\frac{T}{N}\cdot k}\right)\right) + 20tC \cdot N/T} = \frac{\gamma^{\mathrm{cont}}}{4\int_{t=0}^{T} \operatorname{Var}_\pi(f_t)\mathrm{d}t + 10tC},$$

thus the result follows. □

Proof of Theorem 3.4.11. This is similar to the proof of Theorem 3.4.10. We apply Theorem 3.3.7 on the Markov chain $\left\{X_{\frac{T}{N}\cdot k}\right\}_{k\geq 1}$, and then take the limit $N \to \infty$. □

Proof of Theorem 3.4.12. Assume, without loss of generality, that $\mathbb{E}_\pi f_t = 0$ for $0 \leq t \leq T$. We have

$$S' = \int_{t=0}^{t_{\mathrm{ps}}} \sum_{j=0}^{T/t_{\mathrm{ps}}} f_{t+jt_{\mathrm{ps}}}(X_{t+jt_{\mathrm{ps}}})\mathrm{d}t.$$

From the proof of Theorem 3.3.9, it follows that

$$\begin{aligned}
&\mathbb{E}\left(\exp\left(\theta \sum_{j=0}^{T/t_{\mathrm{ps}}} f_{t+jt_{\mathrm{ps}}}(X_{t+jt_{\mathrm{ps}}})\right)\right) \\
&\quad \leq \exp\left(\frac{2\sum_{j=0}^{T/t_{\mathrm{ps}}} \operatorname{Var}_\pi f_{t+jt_{\mathrm{ps}}}}{\gamma((\boldsymbol{P}^*)_{t_{\mathrm{ps}}}\boldsymbol{P}_{k_{\mathrm{ps}}})} \cdot \theta^2 \cdot \left(1 - \frac{10\theta}{\gamma((\boldsymbol{P}^*)_{t_{\mathrm{ps}}}\boldsymbol{P}_{t_{\mathrm{ps}}})}\right)^{-1}\right),
\end{aligned}$$

and the result follows by applying Jensen's inequality with appropriate weights. □

Proof of Theorem 3.4.15. This follows by applying Corollary 3.2.11 to $f(X^{(N)})$, then taking the limit in $N \to \infty$ (and using (3.4.32) and (3.4.33)). □

Chapter 4

Mixing and concentration by Ricci curvature[1]

4.1 Introduction

The coarse Ricci curvature of a Markov chain with metric state space (Ω, d), and kernel $P(x, \mathrm{d}z)$ was defined in Ollivier (2009) as

$$\kappa(x,y) = 1 - \frac{W_1(P_x, P_y)}{d(x,y)} \text{ for } x \neq y, \text{ and } \kappa = \inf_{x,y \in \Omega, x \neq y} \kappa(x,y).$$

where P_x denotes the measure $P(x, \mathrm{d}z)$, and W_1 denotes the Wasserstein distance of P_x and P_y.

It is known that for reversible chains, κ gives a lower bound on the spectral gap: $\gamma \geq \kappa$. It can be also used to bound the mixing time of the chain (known as the Bubley-Dyer path coupling method, see Bubley and Dyer (1997)). The name

[1]This chapter is based on the manuscript Paulin (2013).

curvature comes from the fact that it is linked to the geometric definition of Ricci curvature. One of the motivating examples of Ollivier (2009) is the well known Gromov-Lévy theorem, which it recovers (up to a small constant factor).

When considering Lipschitz functions on Ω under the stationary distribution π of the chain, it is possible to prove variance and concentration bounds, with constants depending on $1/\kappa$, the typical step size of the Markov chain, and the Lipschitz coefficient. In addition to this, one can show concentration inequalities for MCMC empirical averages of Lipschitz functions (see Joulin and Ollivier (2010)).

The coarse Ricci curvature approach have been found to give the right order of concentration and spectral bounds in numerous examples. However, there were also cases where it has not succeeded to give bounds of the correct order. One of them is the split-merge walk on partitions (also called the coagulation-fragmentation chain, see Diaconis, Mayer-Wolf, Zeitouni, and Zerner (2004) for references), where $\kappa = \mathcal{O}(1/N^2)$, which is too small, since $\gamma = \mathcal{O}(1/N)$ in this case. In order to extend the coarse Ricci curvature approach to this situation, we define the multi-step coarse Ricci curvature as

$$\kappa_k(x,y) = 1 - \frac{W_1(P_x^k, P_y^k)}{d(x,y)} \text{ for } x \neq y, \text{ and } \kappa_k = \inf_{x,y\in\Omega, x\neq y} \kappa_k(x,y),$$

which is the coarse Ricci curvature of the k step Markov kernel P^k. We extend the spectral and concentration bounds to this case. We show that for reversible chains, for any $k \in \mathbb{N}$, the spectral gap satisfies $\gamma \geq \kappa_k/k$, and concentration inequalities hold with constants depending on $\sum_{k=0}^{\infty}(1-\kappa_k)$. In particular, this allows us to recover bounds of the correct order of magnitude for the split-merge walk on partitions.

Our concentration bounds essentially mean that if we have a Markov chain that

has small step sizes, and it mixes fast in the multi-step coarse Ricci curvature sense, then the stationary distribution is concentrated. Intuitively, it is clear that stationary distributions of such chains cannot have multiple modes, since they could not mix well by just making local moves. Unimodal distributions tend to satisfy some form of concentration, and as we will see, the strength of the concentration (Gaussian, exponential, or polynomial) is related to the tail behaviour of the step sizes.

We propose several approaches to bound κ_k. The first approach is applicable when the mixing time of the chain can be bounded, and the state space is discrete. In this case, we are able to obtain bounds on κ_k for sufficiently large k, which in turn can imply concentration bounds. We illustrate this with an example about the Curie-Weiss model in critical phase. The second approach gives a recursive lower bound on κ_k. If the curvature is positive in most of the state space, and negative in a small part, then in some situations, this recursive bound can show that κ_k becomes positive for sufficiently large k. An example is given about a random walk on a binary cube with a forbidden region.

Now we explain the organisation of this chapter. In Section 4.2, we introduce the main definitions. Section 4.3 contains our results, in particular, new spectral bounds, concentration inequalities, and moment bounds involving the multi-step coarse Ricci curvature. We also state propositions for bounding κ_k. In Section 4.4, we present some applications. Finally, Section 4.5 contains the proofs of our concentration inequalities.

We end the introduction by a few additional remarks about the related literature. The coarse Ricci curvature approach originates from semigroup tools, which have been used previously in the literature to prove concentration inequalities for

Lipschitz functions of random variables distributed according to the stationary distribution of a Markov process (see Ledoux (2001), Section 2.3). These can be used to prove concentration for the Gaussian measure, and more generally, for log-concave densities. For a recent extension of the coarse Ricci curvature to continuous time Markov processes, see Veysseire (2012a), and Veysseire (2012b). Veysseire (2012) obtains concentration bounds in the case when the coarse Ricci curvature is zero. The coarse Ricci curvature have been used previously, but without geometric interpretation, to bound mixing times, known as the Bubley-Dyer path coupling method. In this sense, it has been also extended to consider multiple steps in the Markov chain, in Dyer, Goldberg, Greenhill, Jerrum, and Mitzenmacher (2000), see also Bhamidi, Bresler, and Sly (2011). The coarse Ricci curvature approach was adapted to graphs in Bauer, Jost, and Liu (2011) and Bauer, Horn, Lin, Lippner, Mangoubi, and Yau (2013), and to adaptive MCMC in Pillai and Smith (2013).

There is another popular curvature notion called the Sturm-Lott-Villani curvature (Lott and Villani (2009), Sturm (2006)). Ollivier (2013) gives a visual introduction to various curvature definitions, and compares them on numerous examples. In the case of Riemann manifolds, Milman (2012a) studies the relation of isoperimetric, functional and transportation cost inequalities, and Milman (2012b) generalises the Gromov-Lévy theorem to compact manifolds with negative curvature. This chapter was motivated by some of the problems of the survey Ollivier (2010). Finally, we note that after we have completed this work, Luczak (2008) have been bought to our attention. It considers similar ideas as ours, and obtains concentration and spectral bounds depending on the contraction properties of the measures describing multiple steps in the Markov chain. The approach was further developed in Luczak (2012),

Brightwell and Luczak (2013b) and Brightwell and Luczak (2013a). Our results in this chapter are more precise, since they take into account the typical size of the jump of the Markov chain, as well as the dimension of the state space, which were not considered in the earlier work. In addition, we also show a recursive bound on the multi-step coarse Ricci curvature, which makes our method easier to apply in practice.

4.2 Preliminaries

We will work with stationary, time homogeneous Markov chains $(X_i)_{i\in\mathbb{N}}$ with transition kernel $P(x, \mathrm{d}y)$ taking values in a Polish metric space (Ω, d). We will denote the stationary distribution of the chain by π. The expected value of a function $f : \Omega \to \mathbb{R}$ under π will be denoted by $\mathbb{E}_\pi(f)$. The jump measure when starting from x will be denoted by P_x, that is, $P_x(\mathrm{d}y) = P(x, \mathrm{d}y)$. For $k \geq 0$, the k-step transition kernel will be denoted by $P^k(x, \mathrm{d}y)$ (in particular, $P^0(x, \mathrm{d}y) = \delta_x(\mathrm{d}y)$, the Dirac-measure concentrated on x).

4.2.1 Ricci curvature

We define the L^1 transportation distance (Wasserstein distance) of two measures on (Ω, d) as

$$W_1(\mu_1, \mu_2) := \inf_{(X,Y)} \mathbb{E}(d(X, Y)), \tag{4.2.1}$$

where the infimum is taken over all couplings (X, Y) of μ_1 and μ_2 (that is, (X, Y) is a random vector taking values on $\Omega \times \Omega$, whose distribution has marginals μ_1, and μ_2). The following definition is a generalisation of Ollivier's coarse Ricci curvature

(Definition 3 of Ollivier (2009)).

Definition 4.2.1 (Multi-step coarse Ricci curvature)**.** Let (Ω, d) and $P(x, \mathrm{d}y)$ be as above. Then for $k \in \mathbb{N}$, $x, y \in \Omega$, we let

$$\kappa_k(x,y) := 1 - \frac{W_1(P_x^k, P_y^k)}{d(x,y)} \text{ if } x \neq y, \text{ and } \kappa_k(x,y) := 1 \text{ if } x = y, \tag{4.2.2}$$

and define the *multi-step coarse Ricci curvature* as $\kappa_k := \inf_{x,y\in\Omega} \kappa_k(x,y)$.

Remark 4.2.2. For $k = 1$, this is just the usual definition of coarse Ricci curvature, that is, $\kappa = \kappa_1$. It is easy to show that $1 - \kappa_i$ satisfies the inequality

$$1 - \kappa_{k+l} \leq (1 - \kappa_k)(1 - \kappa_l) \quad \text{for } k, l \in \mathbb{N}. \tag{4.2.3}$$

4.2.2 Mixing time and spectral gap

We define the total variational distance of two measures P, Q defined on the same state space $(\Omega, \mathcal{F})$ as

$$d_{\mathrm{TV}}(P,Q) := \sup_{A\in\mathcal{F}} |P(A) - Q(A)|, \tag{4.2.4}$$

which is equivalent to

$$d_{\mathrm{TV}}(P,Q) := \inf_{(X,Y)} \mathbb{P}(X \neq Y), \tag{4.2.5}$$

with the infimum taken over all the couplings (X, Y) of P and Q.

We define the mixing time of a time homogeneous Markov chain with general state space in the following way (similarly to Section 4.5 and 4.6 of Levin, Peres, and Wilmer (2009)).

Definition 4.2.3 (Mixing time)**.** Let $X_1, X_2, X_3, \ldots$ be a time homogeneous Markov chain with transition kernel $P(x, \mathrm{d}y)$, state space Ω (a Polish space), and stationary distribution π. Let us denote

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

Let $L_2(\pi)$ denote the Hilbert space of complex valued measurable functions with domain Ω that are square integrable with respect to π, endowed with the inner product $< f, g >_\pi = \int f g^* \mathrm{d}\pi$, and norm $\|f\|_{2,\pi} := \langle f, f \rangle_\pi^{1/2} = \mathbb{E}_\pi \left(f^2\right)^{1/2}$ (we use the same notation for the induced operator norm). P can be then viewed as a linear operator on $L_2(\pi)$, denoted by $\boldsymbol{P}$, defined as

$$(\boldsymbol{P} f)(x) := \mathbb{E}_{P_x}(f),$$

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 transition kernel $P(x, dy)$, and stationary distibution π, we define the *time reversal* of P as the Markov kernel

$$P^*(x, dy) := \frac{P(y, dx)}{\pi(dx)} \cdot \pi(dy). \tag{4.2.6}$$

Then the linear operator $\boldsymbol{P}^*$ is the adjoint of the linear operator $\boldsymbol{P}$ on $L_2(\pi)$. For a Markov chain with stationary distribution π, we define the *spectrum* of the chain as

$$S_2 := \{\lambda \in \mathbb{C} \setminus 0 : (\lambda \mathbf{I} - \boldsymbol{P})^{-1} \text{ does not exist as a bounded lin. oper. on } L_2(\pi)\}.$$

For reversible chains, S_2 lies on the real line.

Definition 4.2.4 (Spectral gap and pseudo spectral gap)**.** The *spectral gap* for reversible chains is

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

For both reversible, and non-reversible chains, the *absolute spectral gap* is

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

In the reversible case, $\gamma \geq \gamma^*$.

The *pseudo spectral gap* of $\boldsymbol{P}$ (introduced in Paulin (2014)) is

$$\gamma_{\text{ps}} := \max_{k \geq 1} \left\{ \gamma((\boldsymbol{P}^*)^k \boldsymbol{P}^k)/k \right\}, \tag{4.2.7}$$

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

Remark 4.2.5. The pseudo spectral gap is similar to the spectral gap in the sense that it allows to obtain variance and concentration bounds on MCMC empirical averages, for example $\text{Var}_\pi((f(X_1) + \dots f(X_N))/N) \leq 4\text{Var}_\pi(f)/(N\gamma_{\text{ps}})$ (see Paulin (2014), Section 3). Moreover, it is related to the mixing time, $\gamma_{\text{ps}} \leq 1/(2t_{\text{mix}})$, and for chains on finite state spaces, $t_{\text{mix}} \leq (1 + 2\log(2) + \log(1/\pi_{\min}))/\gamma_{\text{ps}}$ (here $\pi_{\min} := \min_{x \in \Omega} \pi(x)$).

4.3 Results

In this section, we will present our results based on the multi-step coarse Ricci curvature. In Section 4.3.1, we present a recursive lower bound for κ_k. Section 4.3.2 states spectral bounds, explain the relation of the multi-step coarse Ricci curvature and spectral properties of the Markov chain, while Section 4.3.3 states bounds on the diameter of the state space. In Section 4.3.4, where we state variance, moment, and concentration bounds for Lipschitz functions of random variables distributed according to the stationary distribution of a Markov chain.

4.3.1 Bounding the multi-step coarse Ricci curvature

Our first proposition, the so called geodesic property is useful to get bounds on κ_k (similarly as in Proposition 19 of Ollivier (2009)).

Proposition 4.3.1. *Suppose that (Ω, d) is ϵ-geodesic in the sense that for any two points $x, y \in \Omega$, there exists an integer n, and a sequence $x_0 = x, x_1, \ldots, x_n = y$ such that $d(x, y) = \sum_{i=0}^{n-1} d(x_i, x_{i+1})$ and $d(x_i, x_{i+1}) \leq \epsilon$ for $0 \leq i \leq n-1$. Let $k \geq 1$, then if $\kappa_k(x, y) \geq \kappa_k$ for any pair of points x, y with $d(x, y) \leq \epsilon$, then $\kappa_k(x, y) \geq \kappa_k$ for any pair of points $x, y \in \Omega$.*

Proof. Apply Proposition 19 of Ollivier (2009) to the Markov kernel P^k. □

The following proposition gives a recursive lower bound on the multi-step Ricci curvature $\kappa_k(x, y)$.

Proposition 4.3.2. *For some $x, y \in \Omega, x \neq y$, let (X, Y) be a coupling of P_x and P_y, then*

$$\kappa_{k+1}(x, y) \geq 1 - \mathbb{E}\left(\frac{d(X, Y)(1 - \kappa_k(X, Y))}{d(x, y)}\right).$$

If (X, Y) satisfies that $\mathbb{E}(d(X, Y)) = W_1(P_x, P_y)$ (that is, the coupling "achieves" the Wasserstein distance), then

$$\kappa_{k+1}(x, y) \geq \kappa(x, y) + \mathbb{E}\left(\frac{\kappa_k(X, Y)d(X, Y)}{d(x, y)}\right).$$

Proof. We are going to construct a coupling $X_{k+1} \sim P_x^{k+1}, Y_{k+1} \sim P_y^{k+1}$ as follows. We start from our coupling (X, Y) of P_x and P_y, and for any $a, b \in \Omega$, define

$$\mathcal{L}(X_{k+1}, Y_{k+1}|X = a, Y = b)$$

as the optimal coupling between P_a^k, P_b^k, achieving $\mathbb{E}(d(X_{k+1}, Y_{k+1})|X = a, Y = b) = W_1(P_a^k, P_b^k)$. Then we have

$$\begin{aligned} W_1(P_x^{k+1}, P_y^{k+1}) &\leq \mathbb{E}(d(X_{k+1}, Y_{k+1})) = \mathbb{E}(\mathbb{E}(d(X_{k+1}, Y_{k+1})|X, Y)) \\ &= \mathbb{E}((1 - \kappa_k(X, Y))d(X, Y)), \end{aligned}$$

and thus

$$\begin{aligned} \kappa_{k+1}(x, y) &= 1 - \frac{W_1(P_x^{k+1}, P_y^{k+1})}{d(x, y)} \geq 1 - \frac{\mathbb{E}((1 - \kappa_k(X, Y))d(X, Y))}{d(x, y)} \\ &= 1 - \frac{\mathbb{E}(d(X, Y))}{d(x, y)} + \mathbb{E}\left(\frac{\kappa_k(X, Y)d(X, Y)}{d(x, y)}\right). \end{aligned}$$

Finally, if (X, Y) is the optimal coupling between P_x, and P_y, then $\mathbb{E}(d(X, Y)) = (1 - \kappa(x, y))d(x, y)$, and the second claim of the proposition follows. □

Suppose that everywhere except in a small part of the state space Ω, $\kappa(x, y) > 0$ for neighbouring x and y. Then this result says that $\kappa_{k+1}(x, y)$ can be lower bounded

by some sort of average of $\kappa_k(x, y)$, and for sufficiently large k, the negative curvature may disappear. In Section 4.4.3, we are going to apply this result to a random walk on the binary cube with a forbidden region.

4.3.2 Spectral bounds

Our first result is a bound on the mixing time.

Proposition 4.3.3 (Relation of mixing time and coarse Ricci curvature)**.** *Let* (Ω, d) *be a metric space, and* $P(x, \mathrm{d}y)$ *a Markov kernel. Suppose that* $\operatorname{diam}(\Omega) < \infty$*, and there is* $d_0 > 0$ *such that for any* $x \neq y$*,* $d(x, y) \geq d_0$*. Then*

$$t_{\mathrm{mix}}(\epsilon) \leq \inf\{k : k \geq 1, 1 - \kappa_k \leq \epsilon d_0 / \operatorname{diam}(\Omega)\}. \tag{4.3.1}$$

Conversely, we have, for any $\epsilon > 0$*,* $k \geq t_{\mathrm{mix}}(\epsilon/2)$*,*

$$\kappa_k \geq 1 - \epsilon \cdot \operatorname{diam}(\Omega)/d_0. \tag{4.3.2}$$

Remark 4.3.4. If $\kappa > 0$, then $1 - \kappa_k \leq (1 - \kappa)^k$, thus

$$t_{\mathrm{mix}}(\epsilon) \leq \left\lceil \frac{\log(\epsilon d_0 / \operatorname{diam}(\Omega))}{\log(1 - \kappa)} \right\rceil,$$

which is the well known Bubley-Dyer path coupling bound. Our bound, however, does not require $\kappa > 0$, thus it is more general.

Proof of Proposition 4.3.3. For two disjoint $x, y \in \Omega$, $k \geq 1$, we have

$$d_{\mathrm{TV}}(P_x^k, P_y^k) \leq \frac{W_1(P_x^k, P_y^k)}{d_0} \leq \frac{\operatorname{diam}(\Omega)(1 - \kappa_k)}{d_0}.$$

Averaging out in y gives

$$d_{\mathrm{TV}}(P_x^k, \pi) \leq \frac{W_1(P_x^k, \pi)}{d_0} \leq \frac{\operatorname{diam}(\Omega)(1-\kappa_k)}{d_0},$$

and this is less than equal to ϵ if $1-\kappa_k \leq \epsilon d_0/\operatorname{diam}(\Omega)$. The proof of (4.3.2), based on Proposition 4.3.1, is left to the reader as exercise. □

Now we give lower bounds on the spectral gap and the pseudo spectral gap.

Proposition 4.3.5 (Relation of spectral gap and coarse Ricci curvature)**.** *For reversible chains, for every $k \geq 1$,*

$$\gamma^* \geq 1 - (1-\kappa_k)^{1/k} \geq \frac{\kappa_k}{k}. \tag{4.3.3}$$

Without assuming reversibility, for every $k \geq 1$,

$$\gamma_{\mathrm{ps}} \geq \frac{1-(1-\kappa_k(P^*))(1-\kappa_k)}{k}, \tag{4.3.4}$$

with $\kappa_k(P^*)$ denoting the kth step coarse Ricci curvature of the time reversal of our Markov kernel, $P^*(x, \mathrm{d}y)$.

Remark 4.3.6. In Section 4.4.1, we are going to use this result to obtain a lower bound for the spectral gap of the split-merge walk on partitions. Another application is given in Section 4.4.2, where we use this proposition to bound the pseudo spectral gap of the systemic scan Glauber dynamics in the high temperature regime.

Proof of Proposition 4.3.5. For reversible chains, by applying Proposition 30 of Ollivier (2009) to P^k, we get that $1-\gamma^*(P^k) \leq 1-\kappa_k$, and (4.3.3) follows by the fact that

$1-\gamma^*(P^k) = (1-\gamma^*)^k$. Similarly, applying Proposition 30 of Ollivier (2009) to the reversible kernel $(P^*)^k P^k$, we get $1-\gamma^*((P^*)^k P^k) = 1-\gamma((P^*)^k P^k) \leq \kappa((P^*)^k P^k)$. Now $1-\kappa((P^*)^k P^k) \leq (1-\kappa_k(P^*))(1-\kappa_k)$, thus (4.3.4) follows. □

4.3.3 Diameter bounds

Our first result in this section is an analogue of Proposition 23 of Ollivier (2009).

Proposition 4.3.7 (L^1 Bonnet-Myers theorem)**.** *For $k \geq 1$, let the* k-step jump length *of the random walk at x be*

$$J_k(x) := \int_y d(x,y) \mathrm{d}P_x^k(y).$$

Suppose that for some $k \geq 1$, $\kappa_k(x,y) > 0$ for every $x, y \in \Omega$. Then for every $x, y \in \Omega$, we have

$$d(x,y) \leq \frac{J_k(x) + J_k(y)}{\kappa_k(x,y)},$$

and in particular,

$$\operatorname{diam}(\Omega) \leq \frac{2\sup_x J_k(x)}{\kappa_k} \leq \frac{2k\sup_x J(x)}{\kappa_k}.$$

Proof. Apply Proposition 23 of Ollivier (2009) to P^k. □

Remark 4.3.8. In Section 4.4.1, we are going to apply this proposition to split-merge walk on partitions, and obtain a bound on diameter of Ω of $\mathcal{O}(N)$.

Similarly, we can generalise Proposition 24 of Ollivier (2009).

Proposition 4.3.9 (Average L^1 Bonnet-Myers theorem)**.** *Suppose that for some $k \geq 1$, $\kappa_k(x,y) > 0$ for all $x, y \in \Omega$. Then for any $x \in \Omega$, we have*

$$\int d(x,y)\mathrm{d}\pi(y) \leq \frac{J_k(x)}{\kappa_k},$$

and thus,

$$\int\int d(x,y)\mathrm{d}\pi(x)\mathrm{d}\pi(y) \leq \frac{2\inf_x J_k(x)}{\kappa_k}.$$

Proof. Apply Proposition 24 of Ollivier (2009) to P^k. □

4.3.4 Concentration bounds

Similarly to the results of Ollivier (2009), our concentration bounds will be based on 3 types of quantities related to the multi-step coarse Ricci curvature, the average step size of the Markov chain, and the dimension of the state space. In order to avoid unnecessary repetitions in the statement of the theorems, we introduce some notations (similarly to Definition 18 of Ollivier (2009)).

Definition 4.3.10. Firstly, we make a few definitions related to the multi-step coarse Ricci curvature. Let us define, for any $x, y \in \Omega$,

$$\kappa_\Sigma^c(x,y) := \sum_{i=0}^{\infty}(1-\kappa_k(x,y)), \text{ let } \kappa_\Sigma^c := \sup_{x,y\in\Omega}\kappa_\Sigma^c(x,y), \text{ and } M := \sup_{k\geq 0}(1-\kappa_k).$$

The letter c refers to complement (we add up $1-\kappa_k(x,y)$ instead of $\kappa_k(x,y)$).

Secondly, we state some definitions related to the step size of the Markov chain.

Let the *(coarse) diffusion constant* of the random walk at x be

$$\sigma(x) := \left(\frac{1}{2}\int\int d(y,z)^2 \mathrm{d}P_x(y)\mathrm{d}P_x(z)\right)^{1/2},$$

and let the *average diffusion constant* be

$$\sigma := \left(\int_x \sigma^2(x)\mathrm{d}\pi(x)\right)^{1/2}.$$

Similarly, define the *mean square jump length* as

$$\hat{\sigma}(x) := \left(\int_y d(x,y)^2 \mathrm{d}P_x(y)\right)^{1/2},$$

and the *average mean square jump length* as

$$\hat{\sigma} := \left(\int_x \hat{\sigma}^2(x)\mathrm{d}\pi(x)\right)^{1/2}.$$

Let the *local granularity* be $\sigma_\infty(x) := \frac{1}{2}\mathrm{diam}\,\mathrm{Supp}\,P_x$ (the diameter of the support of P_x), and the *granularity* be $\sigma_\infty := \sup_{x\in\Omega}\sigma_\infty(x)$. Define the *maximal diffusion constant* as $\sigma_{\max} = \sup_{x\in\Omega}\sigma(x)$, and the *maximal mean square jump length* as $\hat{\sigma}_{\max} = \sup_{x\in\Omega}\hat{\sigma}(x)$.

Finally, we state a definition related to the dimension of the state space. Let the *local dimension* at x be

$$n(x) := \frac{\sigma(x)^2}{\sup\{\mathrm{Var}_{P_x} f, f : \mathrm{Supp}\,P_x \to \mathbb{R}\ 1\text{ - Lipschitz}\}}.$$

Remark 4.3.11. Using (4.2.3), we can see that κ_Σ^c can be bounded as

$$\kappa_\Sigma^c \leq \frac{\sum_{i=0}^{k-1}(1-\kappa_i)}{\kappa_k} \leq \frac{kM}{\kappa_k} \text{ for any } k \geq 1. \tag{4.3.5}$$

The random walk can be divided into a drift term (corresponding to the change of the expected location), and a diffusion term (corresponding to the spread in space). The diffusion constant $\sigma^2(x)$ quantifies the diffusion term, when starting from point x.

The local dimension $n(x)$ is a quantity related to the dimension of the state space Ω. In general, when Ω is an N dimensional Euclidean space (or surface of an N dimensional manifold), $n(x)$ is related to N. We always have $n(x) \geq 1$.

Our first concentration result is a variance bound for Lipschitz functions (generalising Proposition 32 of Ollivier (2009)).

Theorem 4.3.12 (Variance bound)**.** *For reversible chains satisfying*

$$\int_y d(x,y)\kappa_\Sigma^c(x,y)\mathrm{d}P_x(y) < \infty \text{ for } \pi - \text{almost every } x, \tag{4.3.6}$$

for any 1-Lipschitz function f on (Ω, d), we have

$$\mathrm{Var}_\pi(f) \leq \int\int \kappa_\Sigma^c(x,y)d(x,y)^2\mathrm{d}\pi(x)\mathrm{d}P_x(y) \leq \frac{1}{2}\kappa_\Sigma^c\hat{\sigma}^2. \tag{4.3.7}$$

More generally, without using reversibility, we have

$$\mathrm{Var}_\pi(f) \leq \left(\sum_{k\geq 0}(1-\kappa_k)^2\right)\mathbb{E}_\pi\left(\frac{\sigma^2}{n}\right). \tag{4.3.8}$$

Our next result is a moment bound for Lipschitz functions of reversible chains.

Theorem 4.3.13 (Moment bound for reversible chains)**.** *For reversible chains satisfying* (4.3.6)*, for any 1-Lipschitz function f on (Ω, d), for any $p \geq 1$, we have*

$$\mathbb{E}_\pi\left([f(X) - \mathbb{E}_\pi(f)]^{2p}\right) \leq \left(\frac{(2p-1)\kappa_\Sigma^c}{2}\right)^p \cdot \mathbb{E}_\pi(\hat{\sigma}^{2p}).$$

Now we state a concentration bound for reversible chains.

Theorem 4.3.14 (Concentration for reversible chains)**.** *For reversible chains satisfying* (4.3.6)*, for any 1-Lipschitz function f on (Ω, d) we have the Gaussian bound*

$$\mathbb{P}_\pi(|f(X) - \mathbb{E}_\pi(f)| \geq t) \leq 2\exp\left(-\frac{t^2}{\kappa_\Sigma^c \cdot \hat{\sigma}_{\max}^2}\right). \tag{4.3.9}$$

For $x \in \Omega$, denote

$$V(x) := \int \kappa_\Sigma^c(x,y) d(x,y)^2 \mathrm{d}P_x(y).$$

Let $L := 4\mathbb{E}_\pi(V)/(\|V\|_{\mathrm{Lip}}\hat{\sigma}^2\kappa_\Sigma^c)$, where $\|V\|_{\mathrm{Lip}}$ is the Lipschitz coefficient of V. Then for any $t \geq 0$,

$$\mathbb{P}_\pi(|f(X) - \mathbb{E}_\pi(f)| \geq t) \leq 2\exp\left(-\frac{t^2}{4\mathbb{E}_\pi(V) + 4L^{-1/2} \cdot t}\right), \tag{4.3.10}$$

More generally, without using reversibility, we have the following concentration bound (generalising Theorem 33 of Ollivier (2009)).

Theorem 4.3.15 (Concentration without reversibility)**.** *For any function f with*

Lipschitz-coefficient $\|f\|_{\mathrm{Lip}}$ *on* (Ω, d)*, let* $S_{\max} := \sup_{x\in\Omega} \frac{\sigma^2(x)}{n(x)}$*, and denote*

$$D_{\max} := 2\|f\|_{\mathrm{Lip}}^2 S_{\max} \cdot \sum_{i=0}^{\infty} \exp\left(\frac{2}{3}(1-\kappa_i)^2\|f\|_{\mathrm{Lip}}^2\right)(1-\kappa_i)^2.$$

Let $t_{\max} := D_{\max}/(6\sigma_\infty)$*, then for* $0 \le t \le t_{\max}$*, we have the Gaussian bound*

$$\mathbb{P}_\pi(|f(X) - \mathbb{E}_\pi(f)| \ge t) \le 2\exp\left(-\frac{t^2}{D_{\max}}\right), \tag{4.3.11}$$

while for $t > t_{\max}$*, we have the exponential bound*

$$\mathbb{P}_\pi(|f(X) - \mathbb{E}_\pi(f)| \ge t) \le 2\exp\left(-\frac{t_{\max}^2}{D_{\max}} - \frac{t - t_{\max}}{3\sigma_\infty}\right). \tag{4.3.12}$$

Theorem 4.3.16. *Alternatively, suppose that* $\sigma^2(x)/n(x) \le S(x)$ *for some* $S : \Omega \to \mathbb{R}$ *(for every* $x \in \Omega$*). Let* K *be a positive integer such that* $\kappa_K > 0$*. Let*

$$D := \|f\|_{\mathrm{Lip}}^2 \mathbb{E}_\pi(S) \cdot \frac{16M^2K}{\kappa_K - \kappa_K^2/4}.$$

Let $\lambda'_{\max} := \min\left(\frac{1}{6M\sigma_\infty\|f\|_{\mathrm{Lip}}}, \frac{\kappa_K}{4KM^2\|S\|_{\mathrm{Lip}}\|f\|_{\mathrm{Lip}}}\right)$*, and* $t'_{\max} := D\lambda'_{\max}/2$*. Then for* $0 \le t \le t'_{\max}$*,*

$$\mathbb{P}_\pi(|f(X) - \mathbb{E}_\pi(f)| \ge t) \le 2\exp\left(-\frac{t^2}{D}\right), \tag{4.3.13}$$

while for $t > t'_{\max}$*,*

$$\mathbb{P}_\pi(|f(X) - \mathbb{E}_\pi(f)| \ge t) \le 2\exp\left(-\frac{t_{\max}^2}{D} - (t - t'_{\max}) \cdot \lambda'_{\max}\right). \tag{4.3.14}$$

Remark 4.3.17. By comparing the concentration inequalities for reversible chains,

and without using reversibility, there are some important differences. Firstly, Theorem 4.3.14 is not using the maximal jump diameter σ_∞, thus it may give better bounds than Theorem 4.3.15 in cases when σ_∞ is very large (or infinity) compared to the typical jump length. However, Theorem 4.3.14 ignores the local dimension $n(x)$, while Theorem 4.3.15 takes it into account, and thus it can give better bounds when $n(x) \gg 1$. The variance, moment, and concentration bounds above can be applied to most of our examples in Section 4.4.

4.4 Applications

In this section, we present some applications of our results. Firstly, in Section 4.4.1, we use the multi-step Ricci curvature (in particular, Proposition 4.3.5 and Theorem 4.3.14) to prove spectral bounds for the transposition walk on the symmetric group, and get concentration inequalities for Lipschitz functions of uniform permutations. In Section 4.4.2 we apply our theorems to Markov chains related to statistical physical models. First, in Section 4.4.2, we show how Dobrushin's interdependence matrix is related to the multi-step Ricci curvature, for Glauber dynamics with random scan and systemic scan. In Sections 4.4.2 and 4.4.2, we apply these bounds to the Curie-Weiss and 1D Ising models, respectively. Finally, in Section 4.4.3, we present an application of the recursive lower bound for κ_k to a random walk on a binary cube with a forbidden region.

4.4.1 Split-merge random walk on partitions

The partitions of N are m-tuples of positive integers $(a_1, \ldots, a_m)$, such that $a_1 \geq a_2 \geq \ldots \geq a_m$, $\sum_{i=1}^{m} a_i = N$, and $m \leq n$. Let us denote the set of the partitions of N by Ω. The split-merge random walk can be thought as the projection of the transposition random walk on the symmetric group S^N to the partitions of N, according to the cycle structure of the permutations. The split-merge walk is defined as in Definition 2 of Bormashenko (2011), as follows.

Assume that we are in $(a_1, \ldots, a_m)$. Then in the following step, we may

1. *Split* – a_i is replaced by $(r, a_i - r)$, with probability a_i/n^2 for every $1 \leq i \leq m, 1 \leq r \leq a_i - 1$.

2. *Merge* – Replace a_i and a_j with $a_i + a_j$, with probability $2a_i a_j/n^2$, for every $1 \leq i < j \leq m$.

3. *Stay* – stay in place with probability $1/n$.

For $x, y \in \Omega$, we define the distance $d(x, y)$ as the minimal number of splits or merges required to get from x to y (or vice-versa). The following proposition estimates the multi-step Ricci curvature κ_k for this random walk on the metric space (Ω, d).

Proposition 4.4.1 (Ricci curvature for the split-merge walk on partitions)**.** *For the split-merge walk on partitions of N, $\kappa > 0$, and thus $\kappa_i > 0$ for any $i \geq 1$. Moreover, there exists $\alpha > 0$, $0 < \beta < 1$ universal constants such that for $k \geq (\alpha + 1/2)N$, $\kappa_k \geq \beta$.*

Proof. First, we are going to show that $\kappa > 0$. By Proposition 4.3.1, it is sufficient to show that

$$W_1(P_x, P_y) \leq (1 - \kappa) d(x, y), \tag{4.4.1}$$

for neighbouring x and y, that is, when $d(x,y) = 1$. Now it is easy to construct a coupling (X,Y) of P_x and P_y such that $d(X,Y) \leq 1$, and $\mathbb{P}(X = Y) = 2/n^2$. This means that (4.4.1) holds with $\kappa = 2/n^2$. The fact that $\kappa_k \geq \beta$ for $k \geq (\alpha + 1/2)N$ follows from Lemma 17 of Bormashenko (2011). □

Now we can apply our results on this example. Firstly, using Proposition 4.3.3, and the facts that $\operatorname{diam}(\Omega) = N - 1$, $d_0 = 1$, and $1 - \kappa_{(\alpha+1/2)N \cdot l} \leq (1-\beta)^l$ for $l \in \mathbb{N}$, we have

$$t_{\text{mix}}(\epsilon) \leq (\alpha + 1/2)N \cdot \frac{\log((N-1)/\epsilon)}{\log(1/(1-\beta))} = \mathcal{O}(N \log(N)).$$

Similarly, using Proposition 4.3.5, we can see that $\gamma^* \geq \frac{\kappa_k}{k} \geq \frac{\beta}{(\alpha+1/2)N} = \mathcal{O}(1/N)$. These are likely to be of the correct order of magnitude, since similar results hold for the transposition walk on the symmetric group (as shown in Diaconis and Shahshahani (1981)). Such bounds could not have been deduced using original coarse Ricci curvature approach of Ollivier (2009), since $\kappa = \mathcal{O}(1/N^2)$.

Applying Proposition 4.3.7 shows that $\operatorname{diam}(\Omega) \leq 2(\alpha + 1/2)N/\beta$, which is the correct order of magnitude.

Finally, in our concentration bounds for reversible chains (Theorem 4.3.12 and Theorem 4.3.14), we have $\kappa_\Sigma^c \leq (\alpha + 1/2)N/\beta$, thus for any $f : \Omega \to \mathbb{R}$ that is 1-Lipschitz with respect to d, $\operatorname{Var}_\pi(f) \leq (\alpha + 1/2)N/\beta$ and

$$\mathbb{P}_\pi(|f(X) - \mathbb{E}_\pi f| \geq t) \leq \exp\left(-\frac{t^2}{(\alpha + 1/2)N/\beta}\right).$$

Note that this result also follows from the concentration result for functions of random permutations (see Maurey (1979), and Talagrand (1995)), since the $d(x,y)$ can be bounded from above by the transposition distance.

It would be interesting to prove similar bounds for the transposition walk on the symmetric group, too. In fact, Bormashenko (2011) uses a connection between the two walks to bound the mixing time of the transposition walk on the symmetric group, based on a coupling argument for the split-merge walk on partitions. However, this approach does not seem to be applicable to the multi-step coarse Ricci curvature.

4.4.2 Glauber dynamics on statistical physical models

In this section, we are going to estimate the coarse Ricci curvature of the Glauber dynamics (with random, and systemic scan) on statistical physical models. A common property of these models is that we have some random variables (spins) $X_1, X_2, \ldots, X_N$, that are dependent on each other, and the strength of their dependence is influenced by a parameter β (inverse temperature).

In the following, in Section 4.4.2, first we define the Dobrushin interdependence matrix (a way to measure the strength of dependence between the random variables), and then state propositions that estimate κ_k in terms of this matrix in the case of Glauber dynamics. In Sections 4.4.2 and 4.4.2, we apply our results to the Curie-Weiss, and the one dimensional Ising models.

Bounds using the Dobrushin interdependence matrix

The following definition originates from Dobrusin (1968) and Dobrushin (1970).

Definition 4.4.2 (Dobrushin interdependence matrix)**.** Let (Λ, d_Λ) be a Polish metric space (of a single spin). Define $\Omega := \Lambda^N$, and for $x, y \in \Omega$, define $d(x, y) = \sum_{i=1}^N d_\Lambda(x_i, y_i)$, where x_i denotes coordinate i of x.

For $x \in \Omega$, denote $x_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N)$. Given a Ω valued random

vector $X = (X_1, \ldots, X_N)$ with distribution μ, we say that a matrix $A := (a_{ij})_{i,j \le N}$ is its *Dobrushin interdependence matrix* if $a_{ii} = 0$ for $i \le N$, and for any $x, y \in \Omega$,

$$W_1(\mu_i(\cdot|x_{-i}), \mu_i(\cdot|y_{-i})) \le \sum_{j=1}^{n} a_{i,j} d_\Lambda(x_j, y_j). \tag{4.4.2}$$

Here $\mu_i(\cdot|x_{-i})$ denotes the conditional distribution of the X_i given $X_{-i} = x_{-i}$, and W_1 denotes the Wasserstein distance with respect to the distance d_Λ. Finally, we say that μ satisfies the *Dobrushin condition* if $\|A\|_1 < 1$.

Remark 4.4.3. A frequently used special case of this is when $d_\Lambda(x_i, y_i) = \mathbb{1}[x_i \neq y_i]$, then $W_1(\mu_i(\cdot|x_{-i}), \mu_i(\cdot|y_{-i}))$ corresponds to the total variational distance. For examples using other types of distances, see Wu (2006).

Proposition 4.4.4 (Glauber dynamics with random scan)**.** *Let (Ω, d), μ and X and A be as in Definition 4.4.2. Consider the Glauber dynamics Markov chain on Ω as follows. In each step, we choose a coordinate I uniformly from $[N]$, and then replace X_I with a conditionally independent copy, given X_{-I}. Then for this Markov chain, we have*

$$\kappa_k \ge 1 - \left\| \left(\frac{N-1}{N} \boldsymbol{I} + \frac{1}{N} \boldsymbol{A} \right)^k \right\|_1. \tag{4.4.3}$$

This implies, in particular, that the absolute spectral gap γ^* satisfies

$$\gamma^* \ge 1 - \mathrm{sp}\left(\frac{N-1}{N} \boldsymbol{I} + \frac{1}{N} \boldsymbol{A} \right) \ge \frac{1 - \mathrm{sp}(\boldsymbol{A})}{N}, \tag{4.4.4}$$

where $\mathrm{sp}(\boldsymbol{A})$ denotes the spectral radius of $\boldsymbol{A}$.

Remark 4.4.5. Notice that $\left\| \left(\frac{N-1}{N} \boldsymbol{I} + \frac{1}{N} \boldsymbol{A} \right)^k \right\|_1$ tends to 0 as $k \to \infty$ if and only if the spectral radius of A is strictly smaller than 1. This follows from the Gelfand's

formula, which says that the spectral radius of a matrix M equals $\lim_{k\to\infty} \|M^k\|^{1/k}$, for any induced matrix norm. This is a less restrictive criteria than $\|A\|_1 < 1$. In particular, $\|A\|_\infty < 1$, or $\|A\|_2 < 1$ also suffices. See Wu (2006) for a spectral gap bound for Markov processes that is similar to (4.4.4).

Proof. The proof is similar to the proof of Theorem 4.3 of Chatterjee (2005). We start by defining a coupling of Ω valued random variables $(X^k, Y^k)_{k\in\mathbb{N}}$, satisfying that $X^k \sim P_x^k, Y^k \sim P_y^k$. First, let $X^0 = x$, and $Y^0 = y$. Suppose that we have already defined $(X^k, Y^k)_{0\le k\le r}$. Then let I_r be uniformly distributed in $[N]$. Now we define X^r and Y^r as equal to X^{r-1} and Y^{r-1} except in their I_rth component. We define $X_{I_r}^r$ and $Y_{I_r}^r$ as the coupling that minimises the Wasserstein distance of the distributions $\mu_{I_r}(\cdot|X_{-I_r}^{r-1}), \mu_{I_r}(\cdot|Y_{-I_r}^{r-1})$ (if the minimising distribution does not exist, then we can make a limiting argument). For this coupling, define the vectors $(\boldsymbol{l}^k)_{k\ge 0}$ taking values in $\mathbb{R}^N$ as $\boldsymbol{l}_i^k := \mathbb{E}(d_\Lambda(X_i^k, Y_i^k))$. Using the definition of the Dobrushin interdependence matrix, we can show that for $k \ge 0$,

$$\boldsymbol{l}^{k+1} \le \left(\frac{N-1}{N}\boldsymbol{I} + \frac{1}{N}\boldsymbol{A}\right)\boldsymbol{l}^k,$$

where the inequality is meant in each component. From this, we can see that

$$\boldsymbol{l}^k \le \left(\frac{N-1}{N}\boldsymbol{I} + \frac{1}{N}\boldsymbol{A}\right)^k \boldsymbol{l}^0,$$

which implies that $1-\kappa_k \le \left\|\left(\frac{N-1}{N}\boldsymbol{I} + \frac{1}{N}\boldsymbol{A}\right)^k\right\|_1$. Finally, (4.4.4) follows from Gelfand's formula, and (4.3.3). □

Proposition 4.4.6 (Glauber dynamics with systemic scan)**.** *Let Ω, μ, X, and A be as in Definition 4.4.2. Consider a Markov chain such that in each step, we go through*

$X_1, \dots, X_n$ in a row, and replace them with a conditionally independent copy given the rest. For $1 \le i \le N$, define B_i as a matrix equal to the identity matrix, except its ith row, which is the same as the ith row of A. Let $B = B_n \cdot B_{n-1} \cdot \ldots \cdot B_1$. Then for $k \ge 1$,

$$\kappa_k \ge 1 - \|B^k\|_1.$$

Remark 4.4.7. Similarly to the random scan case, $\|B^k\|_1 \to 0$ as $k \to \infty$ if and only if the spectral radius of B is less than 1.

Remark 4.4.8. Dyer, Goldberg, and Jerrum (2008) contains an estimation of the mixing time of the systemic scan Glauber dynamics under various forms of the Dobrushin condition. In particular, in Section 7 it is proven that for any $x, y \in \Omega$,

$$d_{\mathrm{TV}}(P_x^k, P_y^k) \le N\|A\|_1^k,$$

implying that

$$t_{\mathrm{mix}}(\epsilon) \le 1 + \frac{\log(N) + \log(1/\epsilon)}{\log(1/\|A\|_1)} \le 1 + \frac{\log(N) + \log(1/\epsilon)}{1 - \|A\|_1}.$$

Proof of Proposition 4.4.6. The proof is similar to the proof of Proposition 4.4.4, but this time we need to show that $\boldsymbol{l}^{k+1} \le \boldsymbol{B}\boldsymbol{l}^k$. The details are left to the reader. □

Curie-Weiss model

Let $\Lambda := \{-1, 1\}$, $\Omega = \Lambda^N$. The natural distance on Λ is $d_\Lambda(a, b) := \mathbb{1}[a \ne b]$, which induces the *Hamming distance* $d(x, y) := \sum_{i=1}^n \mathbb{1}[x_i \ne y_i]$ for $x, y \in \Omega$. For any $\omega \in \Omega$,

let the Hamiltonian function be

$$H_{CW}^{\beta,h}(\omega) := \frac{\beta}{N} \cdot \sum_{1\le i<j\le N} \omega_i\omega_j + h\sum_{i\le N}\omega_i.$$

Here $\beta > 0$ is called the inverse temperature, and h is the external field. Define the probability distribution on Ω as

$$\pi_{CW}^{\beta,h}(\omega) = \exp(H_{CW}^{\beta,h}(\omega))/Z_{CW}^{\beta,h}, \tag{4.4.5}$$

where $Z_{CW}^{\beta,h}$ is a normalising constant. In the zero magnetisation case ($h = 0$), this model is known to undergo phase transition at $\beta = 1$. We call $\beta < 1$ the high-temperature phase, $\beta = 1$ the critical phase, and $\beta > 1$ the low-temperature phase.

When applying the Glauber dynamics chains (with random, or systemic scan) of the previous section to this model (see Propositions 4.4.4 and 4.4.6), the distribution $\pi_{CW}^{\beta,h}$ arises as their stationary distribution. The following proposition estimates the multi-step coarse Ricci curvature of these chains.

Proposition 4.4.9 (Ricci curvature for the Curie-Weiss model)**.** *For the Curie-Weiss model described above, for any h and β, for any $k \ge 2$, we have*

$$\kappa^{Gl.rand.scan.} \ge \left(1 - \beta\frac{N-1}{N}\right)\frac{1}{N},\ \ \kappa^{Gl.sys.scan.} \ge 2 - e^{\beta},$$
$$\kappa_k^{Gl.sys.scan.} \ge 1 - \beta e^{\beta}\left(\beta\frac{N-1}{N}\right)^{k-1},\ \ \gamma_{\mathrm{ps}}^{Gl.sys.scan.} \ge \frac{1-\beta\cdot(N-1)/N}{4}.$$

Finally, for $\beta = 1$ and $h = 0$ (the critical phase), there exists a universal constant $C > 0$ such that for any N, any $k \ge CN^{3/2}\log(N)$, $\kappa_k^{Gl.rand.scan.} \ge 1/2$, and $(\kappa_\Sigma^c)^{Gl.rand.scan.} \le 2CN^{3/2}\log(N)$.

Proof. A simple calculation shows that for the Curie-Weiss model, the following matrix is a Dobrushin interdependence matrix for any β and h (albeit not the sharp one for $h \neq 0$).

$$A^{CW} := \begin{pmatrix} 0 & \beta/N & \beta/N & \beta/N & \dots \\ \beta/N & 0 & \beta/N & \beta/N & \dots \\ \vdots & \vdots & \vdots & \vdots & \dots \\ \beta/N & \beta/N & \beta/N & \dots & 0 \end{pmatrix}.$$

Since $\|A^{CW}\|_1 < \beta(N-1)/N$, $\kappa^{Gl.rand.scan.} \geq 1 - \beta(N-1)/N$ by Proposition 4.4.4. For the Glauber dynamics with systemic scan, we apply Proposition 4.4.6 with the Dobrushin interdependence matrix A^{CW}. Let $x := \beta/N$, then after some calculations, we obtain that the matrix B is given by

$$\begin{aligned} b_{i,1} = 0, \;\; b_{i,i+1} = b_{i,i+2} = \ldots = b_{i,N} = x(1+x)^{i-1}, \\ b_{i+k,i} = x \cdot \left((1+x)^{i+k-1} - (1+x)^k \right), \end{aligned}$$

for $1 \leq i \leq N$, $0 \leq k \leq N-i$. Now for any $k \geq 1$, $\|B^k\|_1 = \max(\mathbf{1} \cdot B^k)$ (maximum column sum), with $\mathbf{1}$ denoting a row vector of ones, and max denoting the maximal element of the vector. After a simple calculation, we get that for $1 \leq i \leq N$,

$$(\mathbf{1} \cdot B)_i = (1+x)^N - (1+x)^{N-i+1},$$

which implies that $\|B\|_1 = \max(\mathbf{1} \cdot B) = (1+x)^N - 1 - x = e^\beta - 1 - \beta/N$, thus by Proposition 4.4.6,

$$\kappa^{Gl.sys.scan.} \geq 2 - e^\beta.$$

As we can see, κ is negative for part of the high temperature case ($\beta < 1$). Now we will use the following lemma.

Lemma 4.4.10. *Let $v = (0, 1/(N-1), 2/(N-1), \ldots, 1)$. Then for B defined as above,*

$$(v \cdot B)_i \leq v_i \cdot \left(\beta \cdot \frac{N-1}{N}\right).$$

This lemma can be proven by straightforward calculations, which we omit.

Now it is easy to see that for $1 \leq i \leq N$,

$$(\mathbf{1} \cdot B)_i \leq ((1+x)^N - (1+x)^{N-1}) \cdot (i-1) = \beta(1+x)^{N-1} \cdot \frac{i-1}{N} \leq \beta \frac{N-1}{N} e^{\beta} \frac{i-1}{N-1},$$

and thus by the above lemma, we can conclude that

$$\|B^k\|_1 = \max(\mathbf{1} \cdot B^k) \leq e^{\beta} \left(\beta \frac{N-1}{N}\right)^k,$$

which implies that for $k \geq 1$,

$$\kappa_k \geq 1 - \|B^k\|_1 \geq 1 - e^{\beta} \left(\beta \frac{N-1}{N}\right)^k.$$

From this, for $0 \leq \beta \leq 1$, with the choice of $k = \lceil 2/(1 - \beta \cdot (N-1)/N) \rceil$ (using the identity $(1-c)^{(1/c)} \leq (1/e)$ for $c > 0$), we get $\kappa_k \geq 1 - 1/e$. By symmetry $\kappa_k(\boldsymbol{P}^*) = \kappa_k$, so using (4.3.4), we get

$$\gamma_{\mathrm{ps}}^{Gl.sys.scan.} \geq (1 - 1/e^2)/\lceil 2/(1 - \beta \cdot (N-1)/N) \rceil \geq \frac{1 - \beta \cdot (N-1)/N}{4}.$$

Finally, we move to the case of the critical phase ($\beta = 1, h = 0$). Theorem 2 of Levin,

Luczak, and Peres (2010) (see also Ding, Lubetzky, and Peres (2009)) shows that the mixing time satisfies $t_{\text{mix}} = \mathcal{O}(N^{3/2})$, thus (4.3.2) gives us the bound on κ_k, and by (4.3.5), we get the bound on κ_Σ^c. □

Substituting the bound $(\kappa_\Sigma^c)^{Gl.rand.scan.} \leq 2CN^{3/2}\log(N)$ and $\hat{\sigma}_{\max} = 1$ to Theorem 4.3.14 leads to the following concentration inequality (a new result).

Proposition 4.4.11. *In the critical phase of the Curie-Weiss model ($\beta = 1, h = 0$), for any $f : \Omega \to \mathbb{R}$ that is 1-Lipschitz with respect to d (Hamming distance), for any $t \geq 0$,*

$$\mathbb{P}(|f(X) - \mathbb{E}f| \geq t) \leq 2\exp\left(-\frac{t^2}{2CN^{3/2}\log(N)}\right),$$

where $X \sim \pi_{CW}^{1,0}$, and C is an universal constant.

Remark 4.4.12. This most likely holds without the $\log(N)$ term as well. The constant in the exponent should be at least of order $N^{3/2}$, as one can see from the limiting distribution of the magnetisation ($f(\omega) = \sum_{i=1}^N \omega_i$), where one has to normalise by $N^{3/4}$ (see Chatterjee and Shao (2011), page 466). Proposition 4 of Chatterjee and Dey (2010) shows a subgaussian ($\exp(-ct^4)$) concentration bound for the magnetisation.

1D Ising model

Let $\Omega = \{-1, 1\}^N$. Let d be the Hamming distance on Ω, as in the previous section. For any $\omega \in \Omega$, let the *Hamiltonian* function be

$$H_{\text{I1D}}^{\beta,h}(\omega) := \frac{\beta}{N} \cdot \sum_{1 \leq i < N} \omega_i\omega_{i+1} + h\sum_{i \leq N} \omega_i.$$

Here $\beta > 0$ is called the inverse temperature, and h is the external field. Define the probability distribution on Ω as

$$\pi_{1D}^{\beta,h}(\omega) = \exp(H_{\mathrm{I1D}}^{\beta,h}(\omega))/Z_{\mathrm{I1D}}^{\beta,h}, \tag{4.4.6}$$

where $Z_{\mathrm{I1D}}^{\beta,h}$ is a normalising constant. This model is known to have no phase transition. The following proposition applies our results on this model, assuming that $h = 0$.

Proposition 4.4.13 (Ricci curvature for 1D Ising model)**.** *For the 1D Ising model described above, for $h = 0$, for any $\beta > 0$, let $\rho := 1/(1 + e^{-4\beta})$, then*

$$\kappa^{Gl.rand.scan.} \geq \frac{2}{N}(1-\rho), \;\; \kappa^{Gl.sys.scan.} \geq 2(1-\rho)/(3/2-\rho),$$
$$\gamma_{\mathrm{ps}}^{Gl.sys.scan.} \geq 2(1-\rho)/(3/2-\rho)^2.$$

Proof. For the 1 dimensional Ising model, the probability of a spin being 1, given that m of it's neighbours are 1, $m = 0, 1, 2$, is

$$\frac{1}{1+\exp(4\beta-2h)}, \frac{1}{1+\exp(-2h)}, \frac{1}{1+\exp(-4\beta-2h)}, \text{ respectively.}$$

It follows that for this model, the Dobrushin matrix is tridiagonal, with the diagonal elements being 0. For $h \leq 0$, the above and below-diagonal elements equal $\frac{1}{1+\exp(-4\beta-2h)} - \frac{1}{1+\exp(-2h)}$, while for $h > 0$, they equal $\frac{1}{1+\exp(-2h)} - \frac{1}{1+\exp(4\beta-2h)}$. In the case of zero external field, $h = 0$, the upper and lower diagonal elements equal $\rho - 1/2$, and $\|A^{\mathrm{I1D}}\|_1 = 2\rho - 1 < 1$. Using this, $\kappa^{Gl.rand.scan.} \geq \frac{2}{N}(1-\rho)$ follows by Proposition 4.4.4.

In the systemic scan case, it is easy to see that for $1 \leq j \leq N$, $b_{j1} = 0$, for $1 < r \leq N$, $b_{r-1,r} = \rho - 1/2$, $b_{r,r} = (\rho - 1/2)^2$, and for $r < j \leq N$, $b_{j,r} = (\rho - 1/2)^{2+j-r}$. This impies that $\|B\|_1 \leq (\rho - 1/2)/(1 - \rho + 1/2)$, and $\kappa^{Gl.sys.scan.} \geq 2(1-\rho)/(3/2-\rho)$ follows by Proposition 4.4.6. Finally, by symmetry and using Proposition 4.3.5, we have

$$\gamma_{\mathrm{ps}}^{Gl.sys.scan.} \geq 1 - (1 - \kappa^{Gl.sys.scan.})^2 \geq 2(1-\rho)/(3/2-\rho)^2. \qquad \square$$

4.4.3 Random walk on a binary cube with a forbidden region

Consider a binary cube $\Omega_0 := \{0,1\}^N$. We call the region $F := \{x \in \Omega_0, \sum x_i < R\}$ the forbidden region. Let $\Omega := \Omega_0 \setminus F$. We consider the following random walk (a version of Glauber dynamics) on Ω. If we are in x, then we pick an index I out of $\{1, \ldots, N\}$ uniformly, and

- if $\sum_{i=1}^N x_i > R$, or if $\sum_{i=1}^N x_i = R$ and $x_I = 0$, then x_I is replaced with an independent Bernoulli(1/2) random variable,
- if $\sum_{i=1}^N x_i = R$, and $x_I = 1$, then we do nothing, and stay in x.

The stationary distribution π is the uniform distribution on Ω (the random walk can be shown to be reversible with respect to this distribution). Because of the geodesic property, it is sufficient to look at $\kappa_k(x,y)$ for neighbouring x and y. Because of symmetry, we can denote this by $\kappa_k(j) := \kappa_k(x,y)$ for x such that $\sum_{i=1}^N x_i = j$, y such that $\sum_{i=1}^N y_i = j+1$, and $\sum_{i=1}^N \mathbb{1}[x_i \neq y_i] = 1$ (for $R \leq j \leq N-1$). Initially, we have negative curvature,

$$\kappa_1(R) = \frac{2-R}{2N}, \kappa_1(j) = \frac{1}{N} \text{ for } R < j \leq N-1.$$

From Proposition 4.3.2, we get the recursive bounds

$$\kappa_{k+1}(R) \geq \frac{2-R}{2N} + \frac{N-1+R}{2N} \cdot \kappa_k(R) + \frac{N-R-1}{2N} \cdot \kappa_k(R+1),$$
$$\kappa_{k+1}(j) \geq \frac{1}{N} + \frac{N-1}{2N} \cdot \kappa_k(j) + \frac{j}{2N} \cdot \kappa_k(j-1) + \frac{N-j-1}{2N} \cdot \kappa_k(j+1)$$

for $R < j \leq N-1$. Notice that all the coefficients of $\kappa_k(j)$ in these inequalities are positive. This implies that if we let $\tilde{\kappa}_1(j) := \kappa_1(j)$ for $R \leq j \leq N-1$, and let

$$\tilde{\kappa}_{k+1}(R) := \frac{2-R}{2N} + \frac{N-1+R}{2N} \cdot \tilde{\kappa}_k(R) + \frac{N-R-1}{2N} \cdot \tilde{\kappa}_k(R+1),$$
$$\tilde{\kappa}_{k+1}(j) := \frac{1}{N} + \frac{N-1}{2N} \cdot \tilde{\kappa}_k(j) + \frac{j}{2N} \cdot \tilde{\kappa}_k(j-1) + \frac{N-j-1}{2N} \cdot \tilde{\kappa}_k(j+1)$$

for $R < j \leq N-1$, then $\kappa_k(j) \geq \tilde{\kappa}_k(j)$ for every $k \geq 1$, $R \leq j \leq N-1$, implying that $\tilde{\kappa}_k := \min_{R \leq j \leq N-1} \tilde{\kappa}_k(j) \leq \kappa_k$. It is easy to conduct numerical simulations to see the behaviour of this recursion. The figures below show this for $N = 500$, $R = 100$.

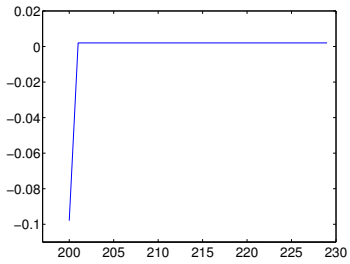
The figures show that initially $\hat{\kappa}_k$ is decreasing, and stays negative, but eventually the positive curvature wins, and $\hat{\kappa}_k$ becomes positive. The following proposition gives bounds on κ_n and κ_Σ^c based on this recursion.

Proposition 4.4.14. *Let* $\rho(R/N) := \frac{1}{2} - \frac{1}{e} - \frac{R}{N-2R}$. *Then for* $R \leq N/10$,

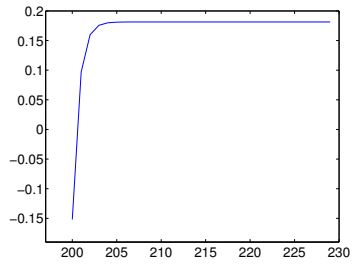
$$\kappa_n \geq \rho(R/N), \text{ and } \kappa_\Sigma^c \leq \frac{N}{\rho(R/N)} \cdot \left(1 + \frac{R}{N-2R}\right).$$

Remark 4.4.15. By Propositions 4.3.3 and 4.3.5, the spectral gap and mixing time of the walk can be bounded as $\gamma \geq \frac{1}{N}\rho(R/N)$, $t_{\text{mix}} \leq 2N\log(N)/\rho(R/N)$. Moreover, by

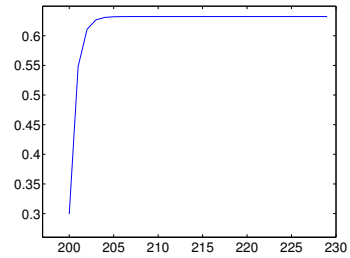
Figure 4.1: Evolution of the multi-step coarse Ricci curvature



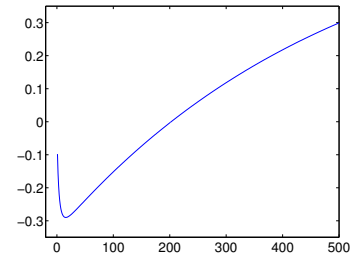
(a) $\tilde{\kappa}_1(j)$ for $R \le j < R + 30$



(b) $\tilde{\kappa}_{100}(j)$ for $R \le j < R + 30$



(c) $\tilde{\kappa}_{500}(j)$ for $R \le j < R + 30$



(d) $\tilde{\kappa}_k$ for $1 \le k \le 500$

Theorem 4.3.14, it follows that for a random vector $X \sim \pi$ and for any 1-Hamming-Lipschitz function f, for any $t \ge 0$,

$$\mathbb{P}(|f(X) - \mathbb{E}(f)| \ge t) \le 2 \exp\left(-\frac{t^2}{N} \cdot \rho(R/N) / \left(1 + \frac{R}{N - 2R}\right)\right).$$

Proof of Proposition 4.4.14. Let $\epsilon := R/N$, and for $0 \le i \le N - R - 1$, $k \ge 1$, let

$$\hat{\kappa}_k(R + i) := -\frac{\epsilon}{1 - 2\epsilon} \cdot \left(\frac{\epsilon}{1 - \epsilon}\right)^i + \left(1 - \exp\left(-\frac{k}{N}\right)\right) - \frac{k - 1}{2N}. \tag{4.4.7}$$

Then it is easy to see that $\hat{\kappa}_1(j) \le \tilde{\kappa}_1(j) \le \kappa_1(j)$ for $R \le j \le N - 1$. Moreover, one

can verify that for every $k \geq 1$,

$$\hat{\kappa}_{k+1}(R) \leq \frac{2-R}{2N} + \frac{N-1+R}{2N} \cdot \hat{\kappa}_k(R) + \frac{N-R-1}{2N} \cdot \hat{\kappa}_k(R+1),$$
$$\hat{\kappa}_{k+1}(j) \leq \frac{1}{N} + \frac{N-1}{2N} \cdot \hat{\kappa}_k(j) + \frac{j}{2N} \cdot \hat{\kappa}_k(j-1) + \frac{N-j-1}{2N} \cdot \hat{\kappa}_k(j+1)$$

for $R < j \leq N-1$, implying that $\hat{\kappa}_k(j) \leq \tilde{\kappa}_k(j) \leq \kappa_k(j)$. The bound on κ_N now follows by noticing that $\kappa_k \geq \hat{\kappa}_k(R)$ for every $k \geq 1$, and the bound on κ_Σ^c follows from (4.3.5). □

4.5 Proofs of concentration results

In this section, we present the proofs of our concentration inequalities. First, we briefly review Chatterjee's method of proving concentration inequalities via Stein's method of exchangeable pairs. We prove our variance and concentration bounds for reversible chains using this approach. Finally, we prove our variance and concentration bounds without using reversibility, by a modification of Ollivier's proofs.

4.5.1 Concentration inequalities via the method of exchangeable pairs

For the proof of our theorems about reversible chains, we will use Stein's method of exchangeable pairs for concentration inequalities, developed in Chatterjee (2005). Let (X, X') be an exchangeable pair taking values in a Polish space Ω. Let $f : \Omega \to \mathbb{R}$, $\mathbb{E}f(X) = 0$, $\mathbb{E}(f(X)^2) < \infty$. Suppose that there is an antisymmetric function

$F : \Omega^2 \to \mathbb{R}$ such that $\mathbb{E}(F(X, X')|X) = f(X)$. Define

$$\Delta(X) := \frac{1}{2}\mathbb{E}(|F(X, X')(f(X) - f(X'))||X), \tag{4.5.1}$$

and assume that $\Delta(X) < \infty$ almost surely. Then the following results hold.

Theorem 4.5.1 (Theorem 3.2 of Chatterjee (2005))**.** *With the above notations,*

$$\mathrm{Var}(f(X)) = \frac{1}{2}\mathbb{E}((f(X) - f(X'))F(X, X')).$$

Theorem 4.5.2 (Theorem 3.14 of Chatterjee (2005))**.** *For any positive integer p, we have*

$$\mathbb{E}\left([f(X) - \mathbb{E}(f)]^{2p}\right) \leq (2p-1)^p \mathbb{E}(\Delta(X)^p).$$

Theorem 4.5.3 (Theorem 3.3 of Chatterjee (2005))**.** *If $\Delta(X) \leq C$ almost surely, then for any $\theta \in \mathbb{R}$, $\mathbb{E}(e^{\theta f(X)}) \leq \exp(\theta^2 C/2)$, and*

$$\mathbb{P}(|f(X) - \mathbb{E}(f(X))| \geq t) \leq 2\exp\left(\frac{-t^2}{2C}\right).$$

Theorem 4.5.4 (Theorem 3.13 of Chatterjee (2005))**.** *Let $r(L) := \frac{\log \mathbb{E}(e^{L\Delta(X)})}{L}$. Then for any $L > 0$ such that $r(L) < \infty$, we have*

$$\mathbb{P}(|f(X) - \mathbb{E}(f(X))| \geq t) \leq 2\exp\left(\frac{-t^2}{2r(L) + 4tL^{-1/2}}\right).$$

Now we show how to find $F(x, y)$ for a given $f(x)$ (based on Section 4 of Chatterjee (2005)). First, notice, that an exchangeable pair (X, X') induces a Markov kernel P,

defined as

$$P(x, A) := \mathbb{P}(X' \in A | X = x) \text{ for every } x \in \Omega, \text{ and every measurable } A \subset \Omega.$$

Conversely, for a reversible Markov kernel P on Ω with stationary distribution π, we define an exchangeable pair as $X \sim \pi$, and $\mathbb{P}(X' \in A | X = x) := P(x, A)$. The following lemma explains the construction of $F(x, y)$ (this is a straightforward extension of Lemma 4.1 of Chatterjee (2005)).

Lemma 4.5.5. *Let X, X' and P as above. Let $f : \Omega \to \mathbb{R}$ be a measurable function with $\mathbb{E}(f(X)) = 0$. Suppose that for every $x, y \in \Omega$, there is a constant $L(x, y) < \infty$ such that $L(y, x) = L(x, y)$,*

$$\sum_{k=0}^{\infty} |P^k f(x) - P^k f(y)| \le L(x, y), \text{ and that } \mathbb{E}(L(X, X')|X) < \infty \text{ almost surely.} \tag{4.5.2}$$

Then the function

$$F(x, y) = \sum_{k=0}^{\infty} (P^k f(x) - P^k f(y)) \tag{4.5.3}$$

satisfies $F(x, y) = -F(y, x)$ and $\mathbb{E}(F(X, X')|X) = f(X)$.

Proof. We have $\mathbb{E}(P^k f(X) - P^k f(X')|X) = P^k f(X) - P^{k+1} f(X)$, and thus

$$\sum_{k=0}^{\infty} \mathbb{E}(P^k f(X) - P^k f(X')|X) = f(X) - P^{N+1} f(X). \tag{4.5.4}$$

Now by (4.5.2), and Lebesgue's dominated convergence theorem, the left hand side will converge to a limit as $N \to \infty$. For the right hand side, we have $P^{N+1} f(y) - P^{N+1} f(x) \to 0$ by (4.5.2) for any $x, y \in \Omega$. The expected value of both sides of

(4.5.4) is 0, so $\lim_{N\to\infty} P^{N+1}f(x) = 0$ for every $x \in \Omega$, and the claim of the lemma follows. □

4.5.2 Concentration of Lipschitz functions under the stationary distribution

We start with the variance bounds.

Proof of Theorem 4.3.12. Without loss of generality, assume $\mathbb{E}f(X) = 0$. Let (X, X') be the exchangeable pair induced by the Markov kernel P, then it is easy to see that

$$|P^k f(x) - P^k f(y)| \le (1 - \kappa_k(x,y))d(x,y),$$

thus (4.5.2) is satisfied with $L(x,y) = d(x,y)\kappa^c_\Sigma(x,y)$. Condition (4.3.6) ensures that $\mathbb{E}(L(X,X')|X) < \infty$ almost surely, and Lemma 4.5.5 gives

$$F(x,y) = \sum_{k=0}^{\infty}(P^k f(x) - P^k f(y)). \tag{4.5.5}$$

By Theorem 4.5.1, we obtain that

$$\begin{aligned}
&\mathrm{Var}(f(X)) = \frac{1}{2}\mathbb{E}((f(X) - f(X'))F(X,X')) = \frac{1}{2}\mathbb{E}\Big[(f(X) - f(X'))\cdot \\
&\cdot \sum_{k=0}^{\infty}(P^k f(X) - P^k f(X'))\Big] \le \frac{1}{2}\mathbb{E}\left(d(X,X')\sum_{k=0}^{\infty}(1 - \kappa_k(x,y))d(X,X')\right) \\
&= \frac{1}{2}\mathbb{E}(\kappa^c_\Sigma(X,X')d(X,X')^2) \le \frac{1}{2}\kappa^c_\Sigma\hat{\sigma}^2.
\end{aligned}$$

Now we turn to the non-reversible case. The proof of this part is similar to the proof of Proposition 32 of Ollivier (2009). Assume first that $\|f\|_\infty < \infty$. Then $\mathrm{Var}(f) < \infty$.

Now we will show that if $\kappa_\Sigma^c < \infty$, then $\mathrm{Var}(P^k f) \to 0$ as $k \to \infty$. Let B_r be a ball of radius r centred at some point in Ω, then we can write

$$\mathrm{Var}(P^k f) = \frac{1}{2}\int\int \left(P^k f(x) - P^k f(y)\right)^2 \mathrm{d}\pi(x)\mathrm{d}\pi(y) \leq 2(1-\kappa_k)^2 r^2 + 2A^2\pi(\Omega \setminus B_r).$$

If we set $r = (1-\kappa_k)^{-1/2}$, then this will tend to 0 as $k \to \infty$, since $\kappa_\Sigma^c < \infty$ implies that $1 - \kappa_k \to 0$. Moreover, if $\kappa_\Sigma^c = \infty$, our bound is vacuous, so there is nothing to prove. Now it is easy to show that

$$\mathrm{Var}(f) = \mathrm{Var}(Pf) + \int_{x\in\Omega} \mathrm{Var}_{P_x}(f)\mathrm{d}\pi(x),$$

and then using $\mathrm{Var}(P^k f) \to 0$, we get

$$\mathrm{Var}(f) = \sum_{k=0}^{\infty}\int_{x\in\Omega} \mathrm{Var}_{P_x}(P^k f)\mathrm{d}\pi(x).$$

Now $\boldsymbol{P}^k(f)$ is $(1-\kappa_k)$-Lipschitz, so by the definition of the local dimension $n(x)$, we have $\mathrm{Var}_{P_x}(\boldsymbol{P}^k(f)) \leq (1-\kappa_k)^2\sigma^2(x)/n(x)$, and (4.3.8) follows. Finally, the $\|f\|_\infty = \infty$ case can be handled by a limiting argument. □

Now we prove concentration for the reversible case.

Proof of Theorem 4.3.14. As in the proof of Theorem 4.3.12, we can show that $F(x,y) = \sum_{k=0}^{\infty}(P^k f(x) - P^k f(y))$, and thus

$$\begin{aligned}\Delta(X) &= \frac{1}{2}\mathbb{E}\left(\left|\sum_{k=0}^{\infty}(P^k f(x) - P^k f(y))\right| \cdot |f(X) - f(X')| \,\middle|\, X\right) \\ &\leq \frac{1}{2}\mathbb{E}\left(\sum_{k=0}^{\infty}(1-\kappa_k)d(X,X')^2 \,\middle|\, X\right) \leq \frac{1}{2}\kappa_\Sigma^c\hat{\sigma}(X)^2 \leq \frac{1}{2}\kappa_\Sigma^c\hat{\sigma}_{\max}^2, \qquad (4.5.6)\end{aligned}$$

and we get (4.3.9) by Theorem 4.5.3. From Theorem 4.5.3 applied to $g(X) = \Delta(X) - \mathbb{E}(\Delta(X))$ it follows that for any $L > 0$,

$$\mathbb{E}(e^{L\Delta(X)}) \le e^{L\mathbb{E}(V(X))} \cdot e^{L^2\|V\|_{\mathrm{Lip}}\kappa_\Sigma^{\mathrm{c}}\hat{\sigma}_{\max}^2/4}.$$

Now choosing L as stated, and applying Theorem 4.5.4 proves (4.3.10). □

Our next proof is the moment bound for reversible chains.

Proof of Theorem 4.3.13. From (4.5.6), we have $\Delta(X) \le \frac{1}{2}\kappa_\Sigma^{\mathrm{c}}\hat{\sigma}(X)^2$, and applying Theorem 4.5.2 leads to this result. □

Now we prove concentration bounds without using reversibility.

The proof of Theorem 4.3.15 is based on the following two lemmas (the first one is a slight variation of Lemma 38 of Ollivier (2009)).

Lemma 4.5.6. *Let $\varphi : \Omega \to \mathbb{R}$ be an α-Lipschitz function. Assume that $\lambda \le 1/(3\sigma_\infty)$. For $r \in \mathbb{R}$, let $g(r) := e^{(2/3)r} \cdot r^2/2$. Then for $x \in \Omega$, we have*

$$(\boldsymbol{P}e^{\lambda\varphi})(x) \le \exp\left(\lambda\boldsymbol{P}\varphi(x) + \lambda^2\frac{\sigma(x)^2}{n(x)} \cdot g(\alpha)\right).$$

Proof. The proof is similar to the original argument, but instead of $\operatorname{diam}\operatorname{Supp} m_x \le 2\sigma_\infty$, now we have $\operatorname{diam}\operatorname{Supp} m_x \le 2\alpha\sigma_\infty$. The details are left to the reader. □

Lemma 4.5.7. *Suppose that a function $f : \Omega \to \mathbb{R}$ satisfies that for $0 \le \lambda \le \lambda_{\max}$,*

$$\mathbb{E}(\exp(\lambda f) \le \exp(\lambda\mathbb{E}(f) + \lambda^2 C).$$

Let $t_{\max} := 2C\lambda_{\max}$*, then for* $0 \le t \le t_{\max}$*, we have*

$$\mathbb{P}(f(X) \ge \mathbb{E}(f) + t) \le \exp\left(-\frac{t^2}{4C}\right),$$

and for $t \ge t_{\max}$*, we have*

$$\mathbb{P}(f(X) \ge \mathbb{E}(f) + t) \le \exp\left(-\frac{t_{\max}^2}{4C} - (t - t_{\max})\lambda_{\max}\right).$$

Proof. This follows by the standard Markov inequality argument. □

Proof of Theorem 4.3.15. Fix some $\lambda \in [0, 1/(3\sigma_\infty)]$. Let $f_0 := f$, and for $k \ge 0$, define f_{k+1} as

$$f_{k+1}(x) := \boldsymbol{P} f_k(x) + \lambda g(\|f_k\|_{\mathrm{Lip}}) \cdot S_{\max}.$$

Lemma 4.5.6 shows that

$$(\boldsymbol{P} f_k)(x) \le e^{\lambda f_{k+1}(x)}, \text{ and thus } (\boldsymbol{P}^k f)(x) \le e^{\lambda f_k(x)}.$$

Since $S_{\max}$ is a constant, we have $\|f_k\|_{\mathrm{Lip}} = \mathrm{Lip}(\boldsymbol{P}^k f) \le (1 - \kappa_k)\|f\|_{\mathrm{Lip}}$, and

$$f_k(x) \le \boldsymbol{P}^k f(x) + \lambda S_{\max} \sum_{i=0}^{k-1} g((1 - \kappa_i)\|f\|_{\mathrm{Lip}}).$$

By taking the limit $k \to \infty$, we get that

$$\lim_{k\to\infty} f_k(x) \le \mathbb{E}_\pi(f) + \frac{\lambda}{4} D_{\max},$$

and thus

$$\mathbb{E}_\pi(e^{\lambda f}) = \lim_{k\to\infty}(\boldsymbol{P}^{\boldsymbol{k}} e^{\lambda f})(x)) \le e^{\lambda \mathbb{E}_\pi(f) + \lambda^2 D_{\max}/4}.$$

We obtain the bounds (4.3.11) and (4.3.12) from Lemma 4.5.7. □

Proof of Theorem 4.3.16. Fix some $\lambda \in [0, \min(1/(3\sigma_\infty), \kappa_K/(2KM\|S\|_{\mathrm{Lip}}))]$. Let $\hat{f}(x) := \hat{f}_0(x) := f(x)/(2M\|f\|_{\mathrm{Lip}})$, and for $k \ge 0$, define $\hat{f}_{k+1}$ as

$$\hat{f}_{k+1}(x) := \boldsymbol{P}\hat{f}_k(x) + \lambda g(\|\hat{f}_k\|_{\mathrm{Lip}}) \cdot S(x).$$

Then Lemma 4.5.6 shows that

$$(\boldsymbol{P}\hat{f}_k)(x) \le e^{\lambda \hat{f}_{k+1}(x)}, \text{ and thus } (\boldsymbol{P}^k \hat{f})(x) \le e^{\lambda \hat{f}_k(x)}.$$

Now for $k \ge 1$, $\hat{f}_k(x)$ as defined above can be expressed as

$$\hat{f}_k(x) = \boldsymbol{P}^k(\hat{f})(x) + \lambda \sum_{i=1}^{k} g(\|\hat{f}_{i-1}\|_{\mathrm{Lip}}) \boldsymbol{P}^{k-i}(S)(x),$$

where

$$\|\hat{f}_k\|_{\mathrm{Lip}} \le (1-\kappa_k)\|\hat{f}\|_{\mathrm{Lip}} + \lambda\|S\|_{\mathrm{Lip}} \sum_{i=1}^{k}(1-\kappa_{k-i}) g(\|\hat{f}_{i-1}\|_{\mathrm{Lip}}). \tag{4.5.7}$$

Now taking the limit $k \to \infty$, we get that

$$\mathbb{E}_\pi(\exp(\lambda\hat{f})) \le \lim_{k\to\infty} \exp(\lambda \hat{f}_k(x)) \le \exp\left(\lambda \mathbb{E}_\pi(\hat{f}) + \lambda^2 \left(\sum_{i=0}^{\infty} g(\|\hat{f}_i\|_{\mathrm{Lip}})\right) \cdot \mathbb{E}_\pi(S)\right). \tag{4.5.8}$$

In order to proceed, we will need to bound $\|\hat{f}_i\|_{\mathrm{Lip}}$ and $\sum_{i=0}^{\infty} g(\|\hat{f}_i\|_{\mathrm{Lip}})$. We claim

that for $\lambda \in [0, \min(1/(3\sigma_\infty), \kappa_K/(2KM\|S\|_{\mathrm{Lip}}))]$, for any $k \in \mathcal{N}$, we have

$$\|\hat{f}_k\|_{\mathrm{Lip}} \leq (1 - \kappa_K/2)^{\lfloor k/K \rfloor}, \tag{4.5.9}$$

and thus

$$\sum_{i=0}^{\infty} g(\|\hat{f}_i\|_{\mathrm{Lip}}) \leq \frac{K}{\kappa_K - \kappa_K^2/4}. \tag{4.5.10}$$

To show this, first note that since $M = \sup_{i \geq 0}(1 - \kappa_i)$, for any $j \geq 0$, we have $1 - \kappa_j \leq M(1 - \kappa_K)^{\lfloor j/K \rfloor}$ (using $(1 - \kappa_{i+j}) \leq (1 - \kappa_i)(1 - \kappa_j)$). This implies that $\sum_{j=0}^{\infty}(1 - \kappa_j) \leq MK/\kappa_K$. Now using the fact that $g(x) \leq x^2$ for $0 \leq x \leq 1$, and the condition $\lambda \in [0, \min(1/(3\sigma_\infty), \kappa_K/(2KM\|S\|_{\mathrm{Lip}}))]$, we can deduce that $\|\hat{f}_j\|_{\mathrm{Lip}} \leq 1$ for any $j \geq 0$. Now let $F_0 := 1$, and for $k \geq 1$, let

$$F_k := \frac{1}{2}(1 - \kappa_K)^{\lfloor k/K \rfloor} + \frac{1}{2}\frac{\kappa_K}{K} \cdot \sum_{i=1}^{k}(1 - \kappa_K)^{\lfloor k - i/K \rfloor} \cdot F_{i-1}^2.$$

Then it follows from (4.5.7) that for any $\lambda \in [0, \min(1/(3\sigma_\infty), \kappa_K/(2KM\|S\|_{\mathrm{Lip}}))]$, any $k \geq 0$ we have $\|f_k\|_{\mathrm{Lip}} \leq F_k$. Now define $G_0 := 1, G_1 := (1 - \kappa_K/2)$, and for $k \geq 2$, let

$$G_k := \frac{1}{2}(1 - \kappa_K)^i + \frac{1}{2}\kappa_K \cdot \left((1 - \kappa_K)^k + \sum_{j=1}^{k-1}(1 - \kappa_K)^j G_{k-1-j}^2 \right).$$

Then it is easy to see that for $k \geq 0$, $F_k \leq G_{\lfloor k/K \rfloor + 1}$, and after some straightforward calculations, we can show that $G_i \leq (1 - \kappa_K/2)^{i-1}$ for any $i \geq 1$. This implies (4.5.9), and by summing up, we get (4.5.10).

Using these two inequalities and (4.5.8), we obtain that for

$$\begin{aligned}&\lambda \in [0, \min(1/(3\sigma_\infty), \kappa_K/(2KM\|S\|_{\mathrm{Lip}}))],\\ &\mathbb{E}_\pi(\exp(\lambda \hat{f})) \leq \exp\left(\lambda \mathbb{E}_\pi(\hat{f}) + \lambda^2 \cdot \frac{K}{\kappa_K - \kappa_K^2/4} \cdot \mathbb{E}_\pi(S)\right),\end{aligned} \tag{4.5.11}$$

which implies that for $\lambda \in \left[0, \min\left(\frac{1}{6M\sigma_\infty\|f\|_{\mathrm{Lip}}}, \frac{\kappa_K}{4KM^2\|S\|_{\mathrm{Lip}}\|f\|_{\mathrm{Lip}}}\right)\right]$, we have

$$\mathbb{E}_\pi(\exp(\lambda f)) \leq \exp\left(\lambda \mathbb{E}_\pi(f) + \lambda^2 \cdot \frac{4M^2\|f\|_{\mathrm{Lip}}^2 K}{\kappa_K - \kappa_K^2/4} \cdot \mathbb{E}_\pi(S)\right). \tag{4.5.12}$$

The tail bounds now follow by Lemma 4.5.7. □

Chapter 5

Convex distance inequality with dependence [1]

5.1 Introduction

The theory of concentration of measure for functions of independent random variables has seen major development since the groundbreaking work of Talagrand (1995) (see the books Ledoux (2001), Dubhashi and Panconesi (2009), and Boucheron, Lugosi, and Massart (2013b)). These inequalities are very useful for obtaining non-asymptotic bounds on various quantities arising from models that are based on collections of independent random variables.

However, for many applications it may be difficult, if not impossible, to describe the model by means of a collection of independent random variables, whereas simpler descriptions based on dependent random variables may be readily available. Such models arise, for example, in statistical physics, where certain distributions can be

[1] This chapter is based on the manuscript Paulin (2014).

described as stationary distributions of appropriate Markov chains. Therefore, it is important to have concentration inequalities that are applicable beyond the independent setting.

In this chapter, we will prove such inequalities for a certain type of dependence, namely for random variables satisfying the so-called the *Dobrushin condition* (however, we believe that the methods presented here can also be adapted to other settings). This condition is satisfied, in particular, in certain statistical physical models when the temperature is sufficiently high, and for sampling without replacement.

Concentration inequalities in the literature for random variables satisfying the Dobrushin condition can be found in the literature (see Külske (2003), Marton (2003), Chatterjee (2005), Djellout, Guillin, and Wu (2004), Wu (2006), Chazottes, Collet, Külske, and Redig (2007), Ollivier (2010), Wang and Wu (2014), Wang (2014)). Most of these results are variants of McDiarmid's bounded differences inequality, only taking into account the maximal deviations

$$\sup_{x_1,\ldots,x_n,x_i'} |f(x_1,\ldots,x_i,\ldots,x_n) - f(x_1,\ldots,x_i',\ldots,x_n)|, \text{ for } 1 \leq i \leq n.$$

In order to get sharper bounds, it is natural to impose stronger conditions on the function f. In this article, we will do this by using the general formalism of (a,b)-*self-bounding functions*, introduced for independent random variables by Boucheron, Lugosi, and Massart (2009).

Our main contribution in this chapter is the following. We will prove concentration inequalities for a slightly restricted subclass of (a,b)-self-bounding functions, which we call (a,b)-$*$-self-bounding (the reason for using the $*$, instead of a letter, is to make it clear that we have two parameters, a and b). We show that our result implies

a version of Talagrand's convex distance inequality for dependent random variables satisfying the Dobrushin condition.

Our approach in this chapter is based on Stein's method of exchangeable pairs, as introduced in Chatterjee (2007). Recently, other variants of Stein's method, size-biasing and zero-biasing, have been adapted to prove concentration inequalities, see Ghosh and Goldstein (2011), and Goldstein and Islak (2013).

It is important to note that for certain types of dependence, such as uniform permutations (Talagrand (1995)) and Markov chains (Marton (1996a), Samson (2000), Marton (2003), and Paulin (2014)) Talagrand's convex distance inequality was shown to hold. However, these approaches do not seem to easily generalise to dependent random variables satisfying the Dobrushin condition.

The rest of this chapter is organised as follows. In Section 5.2, we will introduce the main definitions used in the article. In Section 5.3, we present our main results. In Section 5.4, we discuss three applications, the stochastic salesman problem, the Steiner tree problem, and the total magnetisation of the Curie-Weiss model with external field. In Section 5.5 we prove some preliminary results, and in Section 5.6, we prove our main results. Finally, the Appendix includes a version of Talagrand's convex distance inequality for sampling without replacement.

5.2 Preliminaries

We start by introducing some notation. Let $X := (X_1, \dots, X_n)$ be a vector of random variables, where each X_i takes values in a Polish space Λ_i, and, similarly, let $\Lambda := \Lambda_1 \times \Lambda_2 \times \ldots \times \Lambda_n$, and let $\mathcal{F}$ be the Borel sigma algebra on Λ.

For a vector x in Λ, let $x_{-i} := (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ be the vector created by

dropping the ith coordinate, and set $\Lambda_{-i} := \Lambda_1 \times \ldots \times \Lambda_{i-1} \times \Lambda_{i+1} \times \ldots \times \Lambda_n$. The distribution of the random vector X is denoted by μ, and $(\Lambda, \mathcal{F}, \mu)$ is the probability space induced by X, that is, for $S \in \mathcal{F}$, $\mu(S) = \mathbb{P}(X \in S)$. The marginal distribution of X_i given $X_{-i} = x_{-i}$ will be denoted by $\mu_i(\cdot|x_{-i})$.

We are going to use matrix norms. For an $n \times n$ matrix $A = (a_{ij})_{1 \le i,j \le n}$, we denote its operator norms by $\|A\|_1$, $\|A\|_\infty$ and $\|A\|_2$, respectively. Note that, in particular, $\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^n |a_{ij}|$ and $\|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^n |a_{ij}|$.

Let $g : \Lambda \to \mathbb{R}_+$ be a non-negative function. We will be interested in the concentration properties of $g(X)$. We will denote its centered version by

$$f(x) := g(x) - \mathbb{E}(g(X)).$$

The following definition of self-bounding functions is essentially that of Boucheron, Lugosi, and Massart (2009).

Definition 5.2.1. Let $a, b > 0$. A function $g : \Lambda \to \mathbb{R}_+$ is called (a, b)-*self-bounding* if there exist measurable functions $g_i : \Lambda_{-i} \to \mathbb{R}$, $i = 1, \ldots, n$, such that for every $x \in \Lambda$,

(i) $0 \le g(x) - g_i(x_{-i}) \le 1$ for $1 \le i \le n$, and

(ii) $\sum_{i=1}^n (g(x) - g_i(x_{-i})) \le a g(x) + b$.

A function $g : \Lambda \to \mathbb{R}$ is called *weakly* (a, b)-*self-bounding* if for every $x \in \Lambda$,

(ii') $\sum_{i=1}^n \left(g(x) - g_i(x_{-i})\right)^2 \le a g(x) + b$;

note that (i) is not required in this case.

Remark 5.2.2. If g is (a, b)-self-bounding, then it is also weakly (a, b)-self-bounding.

If g is (a,b)-self-bounding, then we can always take the functions g_i to be

$$g_i(x_{-i}) := \inf_{x_i' \in \Lambda_i} g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n). \tag{5.2.1}$$

We define (a,b)-$*$-self-bounding functions as follows.

Definition 5.2.3. Let $a, b \geq 0$. A function $g : \Lambda \to \mathbb{R}$ is called (a,b)-$*$-*self-bounding* if there exist measurable functions $\alpha_1, \ldots, \alpha_n : \Lambda \to \mathbb{R}$ such that

(i) $0 \leq \alpha_i(x) \leq 1$,

(ii) for every $x, y \in \Lambda$,

$$g(x) - g(y) \leq \sum_{i : x_i \neq y_i} \alpha_i(x),$$

(iii) for every $x \in \Lambda$,

$$\sum_{i=1}^{n} \alpha_i(x) \leq a g(x) + b.$$

Similarly, a function $g : \Lambda \to \mathbb{R}$ is called *weakly* (a,b)-$*$-*self-bounding* if there exists functions $\alpha_1, \ldots, \alpha_n : \Lambda \to \mathbb{R}_+$ such that (ii) above holds, and

(iii') for every $x \in \Lambda$,

$$\sum_{i=1}^{n} \alpha_i(x)^2 \leq a g(x) + b;$$

note that, again, (i) is not required in this case.

Remark 5.2.4. For each $a, b \geq 0$, the following relations hold.

$$\begin{array}{ccc} (a,b)\text{-self-bounding} & \Rightarrow & \text{weakly } (a,b)\text{-self-bounding} \\ \Uparrow & & \Uparrow \\ (a,b)\text{-}*\text{-self-bounding} & \Rightarrow & \text{weakly } (a,b)\text{-}*\text{-self-bounding} \end{array}$$

The reverse implications are false in general.

The following definition allows us to quantify the dependence between the random variables.

Definition 5.2.5 (Dobrushin's interdependence matrix)**.** Suppose $A = (a_{ij})$ is an $n \times n$ matrix with nonnegative entries and zeroes on the diagonal such that for every i, and every $x, y \in \Lambda$,

$$d_{\mathrm{TV}}(\mu_i(\cdot|x_{-i}), \mu_i(\cdot|y_{-i})) \leq \sum_{j \in [n] \setminus \{i\}} a_{ij} \mathbb{1}[x_j \neq y_j], \tag{5.2.2}$$

where d_{TV} denotes the total variational distance (see Section 5.5.1), $\mu_i(\cdot|x_{-i}) = \mathbb{P}(X_i \in \cdot | X_{-i} = x_{-i})$ denotes the marginal of X_i, and $[n] := \{1, \ldots, n\}$. We call such A a *Dobrushin interdependence matrix* for the random vector X (or, equivalently, for the measure μ).

Remark 5.2.6. The condition $\|A\|_1 < 1$ is commonly called the *Dobrushin condition* in the literature. However, some authors use $\|A\|_2 < 1$ or $\|A\|_\infty < 1$ instead. The definition implicitly requires that $\mu_i(\cdot|x_{-i})$ exists for every x_{-i}. This may only be true in some of our applications in an almost sure sense. However, because we are going to assume that our random variables take values in a Polish space, we may use regular conditional probabilities, and change μ on a set of zero probability such that (5.2.2) becomes true everywhere, not just in an almost sure sense (see Faden (1985) for more details on the existence of regular conditional probabilities).

5.3 Main results

In this section, we state our main results regarding concentration for (a, b)-$*$-self-bounding functions, and Talagrand's convex distance inequality. The results apply

to weakly dependent random variables satisfying the Dobrushin condition.

5.3.1 A new concentration inequality for (a, b)-$*$-self-bounding functions

Our main result is a bound on the moment generating function (mgf) of functions of random variables satisfying the Dobrushin condition.

Theorem 5.3.1. *Let $X = (X_1, \ldots, X_n)$ be a vector of random variables, taking values in Λ. Let A be a Dobrushin interdependence matrix for X, and suppose that $\|A\|_1 < 1$ and $\|A\|_\infty \leq 1$. Let $g : \Lambda \to \mathbb{R}$ be a non-negative measurable function such that $g(X)$ has finite mean, denoted by $\mathbb{E}(g)$. Let $a, b \geq 0$.*

1. *If g is (a, b)-$*$-self-bounding, then for $0 \leq \theta \leq (1 - \|A\|_1)/a$,*

$$\log \mathbb{E}\left[e^{\theta(g(X)-\mathbb{E}(g))}\right] \leq \frac{(a\mathbb{E}(g)+b)\theta^2}{2(1-\|A\|_1 - a\theta)}.$$

2. *If g is weakly (a, b)-$*$-self-bounding, then for $0 \leq \theta \leq (1 - \|A\|_1)/(2a)$,*

$$\log \mathbb{E}\left[e^{\theta(g(X)-\mathbb{E}(g))}\right] \leq \frac{(a\mathbb{E}(g)+b)\theta^2}{(1-\|A\|_1 - 2a\theta)}. \tag{5.3.1}$$

3. *Suppose that g is weakly (a, b)-$*$-self-bounding, and in addition, for every $x, x^* \in \Lambda$ differing only in one coordinate, $|g(x) - g(x^*)| \leq 1$. Then for $0 \geq \theta \geq -\frac{1-\|A\|_1}{2a}$, the following inequality holds.*

$$(\log m(\theta))' \geq -\left(e^{-\theta} - 1\right)\frac{2}{1-\|A\|_1}\left(a\mathbb{E}(g) + b - \theta\frac{a(a\mathbb{E}(g)+b)}{2(1-\|A\|_1 + 2a\theta)}\right). \tag{5.3.2}$$

The proof of this is deferred to Section 5.6. As a corollary, we obtain concentration inequalities. For stating them, we will use a constant defined as follows. Let a_c be the unique positive solution of

$$\frac{(\exp(1/4a) - 1)}{1/(4a)} = \frac{8}{5}. \tag{5.3.3}$$

Note that $0.285 < a_c < 0.286$.

Corollary 5.3.2. *Under the conditions of Theorem 5.3.1, we have the following.*

1. *If g is (a, b)-$*$-self-bounding, then for all $t \geq 0$,*

$$\mathbb{P}[g(X) \geq \mathbb{E}(g) + t] \leq \exp\left(-\frac{(1 - \|A\|_1)t^2}{2(a\mathbb{E}(g) + b + at)}\right).$$

2. *If g is weakly (a, b)-$*$-self-bounding, then for all $t \geq 0$,*

$$\mathbb{P}[g(X) \geq \mathbb{E}(g) + t] \leq \exp\left(-\frac{(1 - \|A\|_1)t^2}{4(a\mathbb{E}(g) + b + at)}\right).$$

3. *Suppose that g is weakly (a, b)-$*$-self-bounding, and in addition, for every $x, x^* \in \Lambda$ differing only in one coordinate, $|g(x) - g(x^*)| \leq 1$. If $a \geq a_c(1 - \|A\|_1)$, then for all $t \geq 0$,*

$$\mathbb{P}[g(X) \leq \mathbb{E}(g) - t] \leq \exp\left(-\frac{(1 - \|A\|_1)t^2}{8(a\mathbb{E}(g) + b)}\right),$$

while if $a \leq a_c(1 - \|A\|_1)$, then for all $t \geq 0$,

$$\mathbb{P}[g(X) \leq \mathbb{E}(g) - t] \leq \exp\left(-\frac{t^2}{5(a\mathbb{E}(g) + b)/(1 - \|A\|_1) + (2/3)t}\right).$$

5.3.2 The convex distance inequality for dependent random variables

Recently, Talagrand's convex distance inequality was proven using the weakly self-bounding property in Section 2 of Boucheron, Lugosi, and Massart (2009) (the original proof in Talagrand (1995) was based on mathematical induction). We are going to use similar ideas to prove a version of Talagrand's convex distance inequality based on Theorem 5.3.1 and, hence, applicable to dependent random variables satisfying the Dobrushin condition.

The result is stated in terms of Talagrand's convex distance, which is defined as follows. For $c \in \mathbb{R}^n_+$, and $x, y \in \Lambda$, we define $d_c(x,y) := \sum_{i=1}^n c_i \mathbb{1}[x_i \neq y_i]$. For a point $x \in \Lambda$ and a set $S \subset \Lambda$, we let $d_c(x,S) := \min_{y \in S} d_c(x,y)$ and

$$d_T(x,S) := \sup_{c \in \mathbb{R}^n_+, ||c||_2 = 1} d_c(x,S), \tag{5.3.4}$$

which we call *Talagrand's convex distance* between a point x and a set S.

Theorem 5.3.3. *Let $X := (X_1, \ldots, X_n)$ be a vector of random variables, taking values in a Polish space $\Lambda = \Lambda_1 \times \ldots \times \Lambda_n$, equipped with the Borel σ-algebra $\mathcal{F}$. Let μ be the probability measure on Λ induced by X. Let A be a Dobrushin interdependence matrix for X, and suppose that $\|A\|_1 < 1$ and $\|A\|_\infty \leq 1$. Then for any $S \in \mathcal{F}$,*

$$\mathbb{E}\left[e^{d_T(X,S)^2 \cdot (1 - \|A\|_1)/26.1}\right] \leq \frac{1}{\mu(S)}. \tag{5.3.5}$$

Remark 5.3.4. Inequality (5.3.5) is of the same form as Talagrand's original convex distance inequality in the independent case, but the latter holds with the constant

$(1 - \|A\|_1)/26.1$ being replaced by $1/4$. Our bound takes into account the strength of dependence between the random variables.

The following corollary of the above result generalises the so-called "method of non-uniformly bounded differences" to dependent random variables satisfying the Dobrushin condition.

Corollary 5.3.5. *Let $X = (X_1, \ldots, X_n)$ be a vector of random variables, taking values in Λ, equipped with the Borel σ-algebra $\mathcal{F}$. Let μ be the probability measure on Λ induced by X. Let A be a Dobrushin interdependence matrix for X, and suppose that $\|A\|_1 < 1$ and $\|A\|_\infty \leq 1$. Let $g : \Lambda \to \mathbb{R}$ be a function satisfying that for some positive functions $c_1, \ldots, c_n : \Lambda \to \mathbb{R}_+$,*

$$g(x) - g(y) \leq \sum_{i=1}^{n} c_i(x) \cdot \mathbb{1}[x_i \neq y_i] \tag{5.3.6}$$

for every $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ in Λ, and

$$\sum_{i=1}^{n} c_i^2(x) \leq C \tag{5.3.7}$$

for every x in Λ. Then for any $t \geq 0$,

$$\mathbb{P}(|g(X) - \mathbb{M}(g)| \geq t) \leq 2 \exp\left(\frac{-t^2 \cdot (1 - \|A\|_1)}{26.1C}\right), \tag{5.3.8}$$

where $\mathbb{M}(f)$ denotes the median of $g(X)$ (if the median is not unique, then the result holds for all of them).

Proof. The proof is along the same lines as the proof of Lemma 6.2.1 on page 122 of Steele (1997), except that the constant 4 is replaced by $26.1/(1 - \|A\|_1)$. □

5.4 Applications

In this section, we apply our results to a variant of the stochastic travelling salesmen problem, Steiner trees, the Curie-Weiss model, and exponential random graphs.

5.4.1 Stochastic travelling salesman problem

One important and well studied problem in combinatoric optimisation is the travelling salesman problem (TSP). In the simplest, and most studied case, we are given n points in the unit square $[0,1]^2$, and we are required to find the shortest tour, that is, to find the permutation $\sigma \in S_n$ (S_n denoting the symmetric group) that minimises

$$|x_{\sigma(1)} - x_{\sigma(2)}| + \ldots + |x_{\sigma(n)} - x_{\sigma(1)}|,$$

where $|x - y|$ denotes the Euclidean distance between x and y.

Let us denote the length of the minimal tour by $T(x_1, \ldots, x_n)$. There has been much effort to find efficient algorithms to compute the minimal tour (in general, this is a difficult, NP complete problem, but there are fast algorithms that find a tour that is at most a fixed constant times worse than the optimal tour, see Applegate, Bixby, Chvatal, and Cook (2011) for a recent book on this topic).

From a probabilistic point of view, it is of interest to look at the concentration properties of $T(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ is a random sample from $[0,1]^2$. One of the classical applications of Talagrand's convex distance inequality is to show that, if $X_1, \ldots, X_n$ are i.i.d. uniformly distributed in $[0,1]^2$, then $T(X_1, \ldots, X_n)$ is very sharply concentrated around its median (or equivalently, its expected value), with typical deviations of order 1.

We are going to study a modified version of the travelling salesman problem. Let $\mathcal{A} := \{a_1, \ldots, a_N\}$ be a fixed set of distinct points in $[0,1]^2$. Let $L(x,y) : \mathcal{A}^2 \to \mathbb{R}$ be the *cost function*, satisfying that for some constant $\mathcal{C}$,

$$|x-y| \le L(x,y) \le \mathcal{C}|x-y| \text{ for every } x, y \in \mathcal{A}, \tag{5.4.1}$$

where $|x-y|$ denotes the Euclidean distance of x and y. Note that the cost function does not need to be a metric, and we do not even assume that it is symmetric. A non-symmetric cost function may be used to model the time taken for driving between two locations in a city that are at different elevation, since going uphill can take longer than going downhill.

For any set of distinct points $\{x_1, \ldots, x_n\} \in \mathcal{A}$, we let $T(x_1, \ldots, x_n)$ be the shortest tour through all the points, that is the minimum of the sum

$$L(x(\sigma(1)), x(\sigma(2)) + \ldots + L(x(\sigma(n)), x(\sigma(1)))$$

for $\sigma \in S_n$. Since T is invariant under the permutation of the points, we will also use the notation $T(\{x_1, \ldots, x_n\})$.

Assume that a set of n distinct points are chosen from $\mathcal{A}$ according some distribution μ on all the subsets of size n of $\mathcal{A}$. Let

$$r_{n,1}(\mu) := \sup_{\substack{\mathcal{B} \subset \mathcal{A} \\ |\mathcal{B}| = n-1}} \sup_{b \in \mathcal{A} \setminus \mathcal{B}} \frac{\mu(\mathcal{B} \cup b)}{\sum_{b' \in \mathcal{A} \setminus \mathcal{B}} \mu(\mathcal{B} \cup b')},$$

$$r_{n,2}(\mu) := \sup_{\substack{\mathcal{B}\subset\mathcal{A}\\|\mathcal{B}|=n-2}} \sup_{b,c,d\in\mathcal{A}\setminus\mathcal{B}} \left| \frac{\mu(\mathcal{B}\cup b\cup d)}{\sum_{d'\in\mathcal{A}\setminus(\mathcal{B}\cup b)}\mu(\mathcal{B}\cup b\cup d')} - \frac{\mu(\mathcal{B}\cup c\cup d)}{\sum_{d'\in\mathcal{A}\setminus(\mathcal{B}\cup c)}\mu(\mathcal{B}\cup c\cup d')} \right|,$$

and define the *inhomogeneity coefficient* of this distribution μ as

$$\rho_n(\mu) := n\left(r_{n,1}(\mu) + (N-n)\cdot r_{n,2}(\mu)\right). \tag{5.4.2}$$

This coefficient is related to the distance of the distribution μ from the uniform distribution on all sets of size n, corresponding to sampling without replacement. The following theorem is the main result of this section.

Theorem 5.4.1 (Stochastic TSP for random subsets)**.** *Let $\mathcal{X}$ be a random subset of size n of $\mathcal{A}$, chosen according to a distribution μ, with inhomogeneity coefficient $\rho_n(\mu) < 1$. Then for any $t \geq 0$,*

$$\mu(|T(\mathcal{X}) - \mathcal{M}(T)| \geq t) \leq 4\exp\left(-\frac{t^2(1-\rho_n(\mu))}{1671\mathcal{C}^2}\right), \tag{5.4.3}$$

where $\mathcal{M}(T)$ denotes the median of T.

Remark 5.4.2. The inequality has the same form as the original result in the independent case (in that bound, the exponent is of the form $4\exp(-t^2/64)$).

Example 5.4.3. Now we give a simple example of a distribution μ on $\mathcal{A}$, which we call *weighted sampling without replacement*. Let p be a probability distribution on $[N]$ satisfying that $p(i)$ is strictly positive for every $i \in [N]$. Let us choose a random subset $\mathcal{X} \subset \mathcal{A}$ as follows. Initially, $\mathcal{X}$ is empty. First, we pick an index from $[N]$ according to p, and put the element in $\mathcal{A}$ corresponding this index into $\mathcal{X}$. Then, we pick

another index from $[N]$, according to p conditioned on not choosing the first index. We obtain $\mathcal{X}$ by iterating this procedure n times in total. If we have picked the indices $I_1, \ldots, I_k \in [N]$ in the first k steps, then $\mathbb{P}(k+1\text{th point is } i) = \frac{p(i)}{\sum_{j\in[N]\setminus\{I_1,\ldots,I_k\}} p(j)}$ (for $0 \le k < n$). This means that for any $i_1, \ldots, i_n \in [N]$, we have

$$\begin{aligned}&\mathbb{P}(I_1 = i_1, \ldots, I_n = i_n)\\ &\quad = \mathbb{1}[i_1, \ldots, i_n \text{ are disjoint }] \cdot p(i_1) \cdot \frac{p(i_2)}{\sum_{j\in[N]\setminus\{i_1\}} p_j} \cdot \ldots \cdot \frac{p(i_n)}{\sum_{j\in[N]\setminus\{i_1,\ldots,i_{n-1}\}} p_j}.\end{aligned}$$

Based on this, for a set of n disjoint points $\{a_{i_1}, \ldots, a_{i_n}\} \subset \mathcal{A}$, we define $\mu(\{a_{i_1}, \ldots, a_{i_n}\})$ by averaging over all the possible ways the random variables $I_1, \ldots, I_n$ can take values $i_1, \ldots, i_n$, that is,

$$\mu(\{a_{i_1}, \ldots, a_{i_n}\}) := \frac{1}{n!} \sum_{j_1,\ldots,j_n} \mathbb{P}(I_1 = j_1, \ldots, I_n = j_n),$$

with the summation in $j_1, \ldots, j_n$ is taken over all $n!$ enumerations of $i_1, \ldots, i_n$.

Note that this sampling scheme can be equivalently formulated using independent exponentially distributed random variables with parameters $p_1, \ldots, p_N$ (exponential clocks), where we choose the sets corresponding to the indices of the smallest n such exponential variables (the first n clocks that ring).

Let $p_{\max} := \max_{i\in[N]} p(i)$ and $p_{\min} := \min_{i\in[N]} p(i)$, then an elementary computation shows that for the weighted sampling without replacement scheme,

$$\rho_n(\mu) \le \frac{1}{2}\left(p_{\max}/p_{\min} + (p_{\max}/p_{\min})^2\right) \cdot \frac{n}{N-n}, \tag{5.4.4}$$

which is smaller than 1 if $n < N/\left[1 + \left(p_{\max}/p_{\min} + (p_{\max}/p_{\min})^2\right)/2\right]$.

Sampling without replacement corresponds to the case when $p(i) = 1/N$ for every $i \in [N]$. In this case, the condition of our theorem, $\rho_n(\mu) < 1$, is satisfied if $n < N/2$. In this particular case, using a theorem of Talagrand, we can show that the convex distance inequality holds for any $n \leq N$, which implies that Theorem 5.4.1 also holds for any $n \leq N$. See the Appendix for more details.

Note that it does not seem to be possible to deduce Theorem 5.4.1 using the results of Samson (2000). In the special case when $X_1, \ldots, X_n$ are n samples taken without replacement out of N possibilities, the total variational distance of the distributions $\mathcal{L}(X_l|X_1 = x_1, \ldots, X_k = x_k)$ and $\mathcal{L}(X_l|X_1 = x_1, \ldots, X_{k-1} = x_{k-1}, X_k = x'_k)$ is greater than $1/N$. This means that the above diagonal elements of the mixing matrix are at greater than $1/N$, and the matrix created by taking the square root of every element has L^2 norm of $\mathcal{O}(1+n/\sqrt{N})$. This means that we need to have $n = O(\sqrt{N})$ to obtain concentration results that are only a constant times worse than in the independent case, whereas with our method, this is true for any $n < N/2$.

Now we turn to the proof of Theorem 5.4.1. The proof consists of two parts. Firstly, we compute the coefficients of the Dobrushin interdependence matrix and verify the Dobrushin condition. Secondly, we check that the function T satisfies the conditions of Corollary 5.3.5.

The Dobrushin interdependence matrix is estimated in the following Lemma.

Lemma 5.4.4. *Let μ be a distribution on the subsets of size n of $\mathcal{A}$. Let $X_1, \ldots, X_n$ be random variables taking values in $\mathcal{A}$, distributed as*

$$\mathbb{P}\left(X_1 = a_{i_1}, \ldots, X_n = a_{i_n}\right) = \frac{\mu(\{a_{i_1}, \ldots, a_{i_n}\})}{n!} \text{ for any distinct } i_1, \ldots, i_n \in [N].$$

Then there is a Dobrushin interdependence matrix for $X_1, \ldots, X_n$ such that

$$\|A\|_1, \|A\|_\infty \leq \rho_n(\mu).$$

Proof. Define the event $F_{n-1}(\mathcal{B}, b) := \{\{X_1, \ldots, X_{n-2}\} = \mathcal{B}, X_{n-1} = b\}$ for every $\mathcal{B} \subset \mathcal{A}, |\mathcal{B}| = n-2$ and $b \in \mathcal{A} \setminus \mathcal{B}$. By the definition of the Dobrushin interdependence matrix, using the triangle inequality for the total variational distance, we can set

$$\begin{aligned} a_{n(n-1)} &= \sup_{\substack{\mathcal{B}\subset\mathcal{A},|\mathcal{B}|=n-2,\\ b,c\in\mathcal{A}\setminus\mathcal{B}}} d_{\mathrm{TV}}\big(\mathcal{L}(X_n|F_{n-1}(\mathcal{B},b), \mathcal{L}(X_n|F_{n-1}(\mathcal{B},c))\big) \\ &= \sup_{\substack{\mathcal{B}\subset\mathcal{A},|\mathcal{B}|=n-2,\\ b,c\in\mathcal{A}\setminus\mathcal{B}}} \frac{1}{2} \sum_{d\in\mathcal{A}\setminus\mathcal{B}} \big|\mathbb{P}(X_n = d|F_{n-1}(\mathcal{B},b)) - \mathbb{P}(X_n = d|F_{n-1}(\mathcal{B},c))\big|. \end{aligned}$$

This sum has two type of terms, the first type is when d equals b or c, and the second type is when d equals something else in $\mathcal{A} \setminus \mathcal{B}$. Terms of the first type are less then equal to $r_{n,1}(\mu)$, and terms of the second type are bounded by $r_{n,2}(\mu)$, thus $a_{n(n-1)} \leq \rho_n(\mu)/n$. Because of the symmetry of the distribution of $X_1, \ldots, X_n$, the same holds for every a_{ij}, thus the claim of the lemma follows. □

The following lemma will be used to verify the properties of the function T.

Proposition 5.4.5 (Proposition 11.1 of Dubhashi and Panconesi (2009))**.** *There is a constant $c > 0$ such that, for any set of points $x_1, \ldots, x_n \in [0,1]^2$, there is a permutation $\sigma \in S_n$ satisfying*

$$|x_{\sigma(1)} - x_{\sigma(2)}|^2 + \ldots + |x_{\sigma(n)} - x_{\sigma(1)}|^2 \leq c. \tag{5.4.5}$$

That is, there is a tour going trough all points such that the sum of the squares of the

lengths of all edges in the tour is bounded by an absolute constant c. By the argument outlined in Problem 11.6 of Dubhashi and Panconesi (2009), the above holds with $c = 4$.

The following lemma summarises the properties of the function T required for our proof.

Lemma 5.4.6. *For any* $x, y \in \mathcal{A}^n$, *there are functions* $\alpha_1, \ldots, \alpha_n : [0,1]^2 \to \mathbb{R}_+$ *such that we have*

$$T(x) - T(y) \leq \sum_{i=1}^{n} \alpha_i(x) \mathbb{1}[x_i \neq y_i], \tag{5.4.6}$$

and for any $x \in \mathcal{A}^n$,

$$\sum_{i=1}^{n} \alpha_i^2(x) \leq 64\mathcal{C}^2, \tag{5.4.7}$$

where $\mathcal{C}$ *is as in* (5.4.1).

Proof. For any $x_1, \ldots, x_n \in \mathcal{A}$, let $\hat{\sigma}$ be the permutation in S_n that satisfies (5.4.5). If there are several such permutations, we choose the one that is smallest in the ordering of permutations ranging from $(1, 2, \ldots, n)$ to $(n, n-1, \ldots, 1)$. For $1 \leq i \leq n$, define $\alpha_i(x_1, \ldots, x_n)$ as

$$\alpha_i(x_1, \ldots, x_n) := 2[L(x_{\hat{\sigma}(i-1)}, x_{\hat{\sigma}(i)}) + L(x_{\hat{\sigma}(i)}, x_{\hat{\sigma}(i+1)})],$$

with $i - 1$ and $i + 1$ taken in the modulo n sense. With this choice, inequality (5.4.6) is proven on page 125 of Steele (1997), see also page 144 of Dubhashi and Panconesi (2009). Inequality (5.4.7) follows from Proposition 5.4.5, and the condition $|x - y| \leq L(x, y) \leq \mathcal{C}|x - y|$. □

Now we are ready to prove our concentration result.

Proof of Theorem 5.4.1. The inequality (B.1.3) follows from applying Corollary 5.3.5 to $T(X_1, \ldots, X_n)$, with $\|A\|_1 \leq \rho_n(\mu)$ and $C = 64\mathcal{C}^2$. □

5.4.2 Steiner trees

Suppose that $H = \{x_1, \ldots, x_n\}$ is a set of n distinct points on the unit square $[0, 1]^2$. Then the minimal spanning tree (MST) of H is a connected graph with vertex set H such that the sum of the edge length is minimal (in Euclidean distance). The *minimal Steiner tree* of H is the minimal spanning tree containing H as a subset of its vertices. By the definition, the sum of the edge lengths of this is less than or equal to the sum of the edge lengths of the minimal spanning tree, since we can also add vertices and edges to the graph (an example where they differ is the equilateral triangle, where the minimal Steiner tree adds the centre of mass of the triangle to the graph, thus reducing the total edge length). We denote the sum of the edge lengths of the minimal Steiner tree by $S(x_1, \ldots, x_n)$. Note that this is invariant to permutations of $x_1, \ldots, x_n$, thus we can equivalently denote it by $S(\{x_1, \ldots, x_n\})$.

This is a quantity of great practical importance, since it expresses the minimal amount of interconnect needed between the points $x_1, \ldots, x_n$. It has found numerous applications in circuit and network design. Hwang, Richards, and Winter (1992) is a popular book on this subject.

From a probabilistic perspective, a problem of interest is to quantify the behaviour of $S(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are random variables that are i.i.d. uniformly distributed on $[0, 1]^2$. Steele (1997) has proven that the total length of the minimal Steiner tree, $S(X_1, \ldots, X_n)$, is sharply concentrated around its median, with typical deviations of order 1.

Here we study a modified version of this problem, when we choose a random subset of size n from a set of points $\mathcal{A} := \{a_1, \ldots, a_N\}$ in $[0,1]^2$. Let μ be a probability measure on such subsets, and denote its inhomogeneity coefficient defined in (5.4.2) by $\rho_n(\mu)$. Using our version of Talagrand's convex distance inequality for dependent random variables, we obtain the following concentration bound.

Theorem 5.4.7 (Minimal Steiner tree for random subsets)**.** *Let $\mathcal{X}$ be a random subset of size n of $\mathcal{A}$, chosen according to a distribution μ, with inhomogeneity coefficient $\rho_n(\mu) < 1$. Then for any $t \geq 0$,*

$$\mathbb{P}(|S(\mathcal{X}) - \mathcal{M}(S)| \geq t) \leq 4 \exp\left(-\frac{t^2(1-\rho_n(\mu))}{520000}\right), \tag{5.4.8}$$

where $\mathcal{M}(S)$ denotes the median of S.

The proof consists, again, of two parts. First, we bound the Dobrushin interdependence matrix, then show that the function S satisfies the conditions of our version of the method of non-uniformly bounded differences for dependent random variables (Corollary 5.3.5). The first part is proven in Lemma 5.4.4. For the second part, we are going to use the following lemma.

Lemma 5.4.8 (Steele (1997), page 107, equation (5.26))**.** *Let us denote the edge lengths of the minimum spanning tree for $x_1, \ldots, x_n \in [0,1]^2$ by $e_1, \ldots, e_{n-1}$. Then for some universal constant c,*

$$e_1^2 + \ldots + e_{n-1}^2 \leq c, \tag{5.4.9}$$

in particular, we can choose $c = 410$ (see page 108 of Steele (1997)). If there are multiple minimal spanning trees, then this holds for each of them.

The conditions on S are verified in the following lemma.

Lemma 5.4.9. *For any $x_1, \ldots, x_n \in [0,1]^2$, denote $x = (x_1, \ldots, x_n)$, and for $1 \leq i \leq n$, define $\alpha_i(x)$ as two times the length of the incurring edges in the minimal spanning tree of $x_1, \ldots, x_n$. Then for any $x, y \in ([0,1]^2)^n$, we have*

$$S(x) - S(y) \leq \sum_{i=1}^{n} \alpha_i(x) \cdot \mathbb{1}[x_i \neq y_i].$$

Moreover, for any $x \in ([0,1]^2)^n$,

$$\sum_{i=1}^{n} \alpha_i^2(x) \leq 19680.$$

Proof. The first claim is proven on pages 123-124 of Steele (1997). For the second claim, first notice that the vertices in the minimum spanning tree can have degree at most 6. Now for any 6 reals $z_1, \ldots, z_6$, we have $(z_1 + \ldots + z_6)^2 \leq 6(z_1^2 + \ldots + z_6^2)$, and every edge belongs to two vertex so it is counted twice, thus by Lemma 5.4.8, we have

$$\sum_{i=1}^{n} \alpha_i^2(x) \leq 6 \cdot 2^2 \cdot 2 \sum_{i=1}^{n-1} e_i^2 \leq 19680.$$

□

Now we are ready to prove our concentration result.

Proof of Theorem 5.4.7. Using Lemma 5.4.4 and Lemma 5.4.9, the statement of the theorem follows by applying Corollary 5.3.5 with $\|A\|_1 = \|A\|_\infty = \rho_n(\mu)$ and $C = 19680$. □

5.4.3 Curie-Weiss model

The Curie-Weiss model of ferromagnetic interaction is the following. Consider the state space $\Lambda = \{-1, 1\}^n$, and denote an element of the state space (a configuration) by $\sigma = (\sigma_1, \ldots, \sigma_n)$. Define the Hamiltonian for the system as

$$H(\sigma) := \left(\beta \frac{1}{n} \sum_{1 \le i < j \le n} \sigma_i \sigma_j + h \sum_{i=1}^{n} \sigma_i \right),$$

and the probability density

$$p_\beta(\sigma) := \frac{\exp(\beta H(\sigma))}{Z(\beta, h)},$$

where $Z(\beta, h) := \sum_{\sigma \in \Lambda} \exp(\beta H(\sigma))$ is the normalizing constant. The following proposition gives bounds on the Dobrushin interdepence matrix for this model.

Proposition 5.4.10. *For σ as above, the Dobrushin interdependence matrix A satisfies*

$$\|A\|_1, \|A\|_\infty, \|A\|_2 < \beta.$$

Proof. We will now calculate the Dobrushin interdependence matrix for this system. Suppose first that $h = 0$. Let x and y be two configurations, then we want to bound

$$d_{\mathrm{TV}}(\mu_i(\cdot|x_{-i}), \mu_i(\cdot|y_{-i}))$$

Since σ_i can only take values 1 or -1, so the total variation distance is simply

$$d_{\mathrm{TV}}(\mu_i(\cdot|x_{-i}), \mu_i(\cdot|y_{-i})) = |\mathbb{P}(\sigma_i = 1|x_{-i}) - \mathbb{P}(\sigma_i = 1|y_{-i})|.$$

Now by writing $m_i(x) := \frac{1}{n}\sum_{j:j\neq i} x_j$ and $m_i(y) := \frac{1}{n}\sum_{j:j\neq i} y_j$, we can write

$$\mathbb{P}(\sigma_i = 1|x_{-i}) = \frac{\exp(\beta m_i(x))}{\exp(\beta m_i(x)) + \exp(-\beta m_i(x))},$$

so by denoting

$$r(t) := \frac{\exp(t)}{\exp(t) + \exp(-t)} = \frac{1}{1 + \exp(-2t)}, \tag{5.4.10}$$

we can write

$$|\mathbb{P}(\sigma_i = 1|x_{-i}) - \mathbb{P}(\sigma_i = 1|y_{-i})| = |r(\beta m_i(x)) - r(\beta m_i(y))|.$$

Now it is easy to check that $|r'(t)| \leq \frac{1}{2}$, and changing one spin in x can change m_i at most by $2/n$. From this, we obtain a Dobrushin interdependence matrix A with $a_{ij} = \frac{\beta}{n}$ for $i \neq j$. For this A, it is easy to see that

$$\|A\|_1 = \|A\|_\infty = \|A\|_2 = \beta\left(1 - \frac{1}{n}\right) < \beta. \qquad \square$$

Thus for the high temperature case $0 \leq \beta < 1$, we can apply Corollary 5.3.2 to obtain concentration inequalities.

In the case when writing the conditional probabilities for $h \neq 0$, one can show that in the above argument, $r(t)$ in (5.4.10) gets replaced by $r(t,h) := \frac{\exp(t+h)}{\exp(t+h)+\exp(-t-h)}$. This function still satisfies that $|\frac{\partial}{\partial t} r(t,h)| \leq 1/2$, thus A as defined above is a Dobrushin interdependence matrix in this case as well.

Now we are going to show a concentration inequality for the average magnetization of the Curie-Weiss model. Let us denote the average magnetization by $m := \frac{1}{n}\sum_{i=1}^n \sigma_i$. We have the following proposition.

Proposition 5.4.11. *For the above model, when $0 \leq \beta < 1$, and $h \geq 0$, we have*

$$\mathbb{P}(m(\sigma) \geq \mathbb{E}(m(\sigma)) + t) \leq \exp\left(-\frac{n(1-\beta)t^2}{16(1-\tanh(h)+4/((1-\beta)\sqrt{n})}\right)$$
$$\mathbb{P}(m(\sigma) \leq \mathbb{E}(m(\sigma)) - t) \leq \exp\left(-\frac{n(1-\beta)t^2}{4[1-\tanh(h)+4/((1-\beta)\sqrt{n})]+4t}\right).$$

Remark 5.4.12. Since $1-\tanh(h) \leq 2\exp(-2h)$ for $h \geq 0$, this proposition is better for large values of h than what we could obtain from McDiarmid's bounded differences inequality (Theorem 4.3 of Chatterjee (2005)). That result uses only the Hamming Lipschitz property, and gives bounds of order $\exp(-n(1-\beta)t^2))$, which does not capture the fact that in such cases σ_i and thus $m(\sigma)$ has small variance.

Proof of Proposition 5.4.11 . Let $n_-(\sigma) = \sum_{i=1}^n \mathbb{1}[\sigma_i = -1]$ be the number of -1 spins, then $m = \frac{n-2n_-}{n}$, and for $t \geq 0$,

$$\mathbb{P}(m(\sigma) \geq \mathbb{E}(m(\sigma)) + t) = \mathbb{P}\left(n_-(\sigma) \leq \mathbb{E}(n_-(\sigma)) - \frac{n}{2}t\right), \tag{5.4.11}$$
$$\mathbb{P}(m(\sigma) \leq \mathbb{E}(m(\sigma)) - t) = \mathbb{P}\left(n_-(\sigma) \geq \mathbb{E}(n_-(\sigma)) + \frac{n}{2}t\right). \tag{5.4.12}$$

Here $n_-(\sigma)$ is a sum of non-negative variables, so one can easily see that it is $(1,0)$-$*$-self-bounding, and thus, by Theorem 5.3.1, we have for every $t \geq 0$,

$$\mathbb{P}(n_-(\sigma) \geq \mathbb{E}(n_-(\sigma)) + t) \leq \exp\left(-\frac{(1-\beta)t^2}{2\mathbb{E}(n_-(\sigma)) + 2t}\right) \tag{5.4.13}$$
$$\mathbb{P}(n_-(\sigma) \leq \mathbb{E}(n_-(\sigma)) - t) \leq \exp\left(-\frac{(1-\beta)t^2}{8\mathbb{E}(n_-(\sigma))}\right). \tag{5.4.14}$$

In order to apply this bound, we will need to estimate $\mathbb{E}(n_-(\sigma)) = n(1-\mathbb{E}(m))/2$. For this, we are going to use Proposition 1.3 of Chatterjee (2007), stating that for

any $t \geq 0$,

$$\mathbb{P}\left(m(\sigma) - \tanh(\beta m(\sigma) + h) \geq \frac{\beta}{n} + \frac{t}{\sqrt{n}}\right) \leq \exp(-t^2/(4+4\beta)), \tag{5.4.15}$$

and the same bound holds for the lower tail as well. Here we have replaced βh with h in the equation of Proposition 1.3 because of the different definition of the Hamiltonian of the model. Now for $0 \leq \beta < 1$, the equation $m = \tanh(\beta m + h)$ admits a unique solution in m, which we denote by $m^*(h)$.

For $0 \leq \beta \leq 1$, (5.4.15) can be further bounded by $\exp(-nt^2/8)$, moreover, for any $x \geq 0$, $\mathbb{P}(|m(\sigma) - m^*| \geq x/(1-\beta)) \leq \mathbb{P}(|m(\sigma) - \tanh(\beta m(\sigma) + h)| \geq x)$, and thus for any $t \geq 0$,

$$\mathbb{P}\left((m(\sigma) - m^*) \geq \left(\frac{1}{1-\beta}\right) \cdot \left(\frac{1}{n} + \frac{t}{\sqrt{n}}\right)\right) \leq \exp(-t^2/8),$$

and the same inequality holds for the lower tail as well, but with $m(\sigma) - m^*$ replaced by $m^* - m(\sigma)$. From this, using integration by parts, we obtain that

$$\mathbb{E}((m(\sigma) - m^*)_+), \mathbb{E}((m(\sigma) - m^*)_-) \leq \frac{1}{1-\beta} \cdot \frac{1}{n} + \frac{1}{1-\beta} \cdot \frac{1}{\sqrt{n}} \cdot \sqrt{2\pi} \leq \frac{4}{(1-\beta)\sqrt{n}},$$

implying that $|\mathbb{E}(m(\sigma)) - m^*| \leq 4/((1-\beta)\sqrt{n})$. Now it is easy to see that for $h \geq 0$, we have $m^*(h) \geq \tanh(h)$, and thus $\mathbb{E}(m(\sigma)) \geq \tanh(h) - 4/((1-\beta)\sqrt{n})$ and

$$\mathbb{E}(n_-(\sigma)) \leq n(1 + 4/((1-\beta)\sqrt{n}) - \tanh(h))/2.$$

Now the results follow by combining this with equations (5.4.11), (5.4.12), (5.4.13) and (5.4.14). □

5.4.4 Exponential random graphs

Exponential random graph models are increasingly popular for modelling network data (see Chatterjee and Diaconis (2013)). For a graph with n vertices, the edges are distributed according to a probability distribution of the form

$$p_\beta(G) := \exp\left(\sum_{i=1}^{k} \beta_i T_i(G) - \psi(\beta)\right), \tag{5.4.16}$$

where $\beta = (\beta_1, \ldots, \beta_k)$ is a vector of real parameters, and $T_1, \ldots, T_k$ are functions on the space of the graphs (T_1 is usually the number of edges, while the rest can be the number of triangles, cycles, etc.), and $\psi(\beta)$ is the normalising constant.

The simplest special case of this model is the Erdős-Rényi graph. Let E be the number of edges of the graph, and let $0 < p < 1$ be a parameter, then in this case,

$$p_\beta(G) := p^E(1-p)^{n(n-1)/2-E} = \exp\left(\log\left(\frac{p}{1-p}\right)E + \log(1-p)n(n-1)/2\right).$$

In this case, the edges are i.i.d. random variables distributed according to the Bernoulli distribution with parameter p.

A more complex model, which was analysed in Chatterjee and Diaconis (2013), has the distribution

$$p_{\beta_1,\beta_2}(G) = \exp\left(2\beta_1 E + \frac{6\beta_2}{n}\Delta - n^2\psi_n(\beta_1,\beta_2)\right),$$

where E denotes the number of edges, Δ denotes the number of triangles, and $\psi_n(\beta_1, \beta_2)$ is the normalising constant. Note that in this case, the edges are no longer independent, because the number of triangles introduces a form of dependence into

the model.

In general, for any model of the type (5.4.16), there is a certain set $\mathcal{D} \subset \mathbb{R}^k$ of non-zero volume such that when the parameters $\beta \in \mathcal{D}$, the edges, as random variables, satisfy the Dobrushin condition (that is, there is an interdependence matrix such that $\|A\|_1 < 1$ and $\|A\|_\infty < 1$). This fact can be shown by a simple continuity argument, since the random variables are independent when $\beta = 0$. The set $\mathcal{D}$ is analogous to the high-temperature phase of statistical physical models.

The following theorem, based on our new concentration inequality for (a, b)-*-self-bounding functions, establishes concentration inequalities for subgraph counts in exponential random graph models in the high temperature phase.

Theorem 5.4.13 (Subgraph counts in exponential random graphs)**.**
Let $\Lambda := \{0,1\}^{n(n-1)/2}$, and let $X := (X_{ij})_{1 \le i < j \le n}$ be the edges of an exponential random graph, taking values in Λ, distributed according to p_β, as defined by (5.4.16)*. Suppose that $\beta \in \mathcal{D}$.*

Let S be a fixed graph with n_S vertices and e_S edges. Let N_S denote the number of copies of S in our exponential random graph, then for any $t \ge 0$,

$$\mathbb{P}(N_S - \mathbb{E}(N_S) \ge t) \le \exp\left(\frac{(1 - \|A\|_1)t^2}{2\binom{n-2}{n_S-2} e_S \cdot (\mathbb{E}(N_S) + t)}\right), \tag{5.4.17}$$

$$\mathbb{P}(N_S - \mathbb{E}(N_S) \le -t) \le \exp\left(\frac{(1 - \|A\|_1)t^2}{8\binom{n-2}{n_S-2} e_S \cdot \mathbb{E}(N_S)}\right). \tag{5.4.18}$$

Remark 5.4.14. By the number of copies of S, we mean the number of subsets of size n_S of the set of n vertices of our graph such that the corresponding subgraph contains S. A of similar concentration inequality can be shown to hold for the maximal degree among all the vertices (see Example 6.13 of Boucheron, Lugosi, and Massart (2013b)),

which can be shown to be $(1, 0)$-*-self-bounding. Our results are sharper than what we could obtain using Theorem 4.3 of Chatterjee (2005) (McDiarmid's bound differences inequality for dependent random variables satisfying the Dobrushin condition).

Proof of Theorem 5.4.13. The proof is based on the *-self-bounding property of N_S. If we add an edge to X, then N_S will increase, or stay the same, while if we erase an edge from X, then N_S will decrease, or stay the same. For $x \in \Lambda$, $1 \leq i < j \leq n$, let $\alpha_{i,j}(x)$ be the number of copies of S in x that contain the edge (i, j). Then $0 \leq \alpha_{i,j}(x) \leq \binom{n-2}{n_S-2}$, and we can see that for any $x, y \in \Lambda$,

$$N_S(x) - N_S(y) \leq \sum_{1 \leq i < j \leq n} \alpha_{i,j}(x) \mathbb{1}[x_{ij} \neq y_{ij}].$$

Moreover, since S contains e_S edges, we have

$$\sum_{1 \leq i < j \leq n} \alpha_{i,j}(x) \leq e_S N_S(x).$$

This means that $N_S(x)/\binom{n-2}{n_S-2}$ is $(e_S, 0)$-*-self-bounding, and the results follow by Corollary 5.3.2. □

5.5 Preliminary results

In this section, we will prove some preliminary results needed for proving our main results from Section 5.3. First, we prove a lemma about the total variational distance. After this, review the basics of the concentration inequalities by Stein's method of exchangeable pairs approach. Finally, we prove some lemmas about bounding moment generating functions.

5.5.1 Basic properties of the total variational distance

The total variational distance of two probability distributions μ_1 and μ_2 defined on the same measurable space $(\mathcal{X}, \mathcal{F})$ is defined as

$$d_{\mathrm{TV}}(\mu_1, \mu_2) = \sup_{S \in \mathcal{F}} |\mu_1(S) - \mu_2(S)|. \tag{5.5.1}$$

The following lemma proposes a coupling related to the total variational distance that we are going to use.

Lemma 5.5.1. *Let μ_1 and μ_2 be two probability measures on a Polish space $(\mathcal{X}, \mathcal{F})$. Then for any fixed q with $d_{\mathrm{TV}}(\mu_1, \mu_2) \leq q \leq 1$, we can define a coupling of independent random variables χ, B, C, D such that χ has Bernoulli distribution with parameter q, and the random variables*

$$X := (1 - \chi)B + \chi C, \quad Y := (1 - \chi)B + \chi D \tag{5.5.2}$$

satisfy that $X \sim \mu_1$, $Y \sim \mu_2$.

Proof. The proof is similar to Problem 7.11.16 of Grimmett and Stirzaker (2001). We define the measure $\mu_{12}(\cdot)$ on $(\mathcal{X}, \mathcal{F})$ as $\mu_{12}(S) = \frac{\mu_1(S)+\mu_2(S)}{2}$. Then μ_1 and μ_2 are both absolutely continuous with respect to μ_{12}, thus we can define the Radon-Nikodym derivatives $f(x) := \frac{d\mu_1}{d\mu_{12}}(x)$ and $g(x) := \frac{d\mu_2}{d\mu_{12}}(x)$ for almost every $x \in \Omega$.

The density of random variables B, C and D with respect to μ_{12} can be defined in terms of $f(x)$ and $g(x)$ as follows. Let us define $h : \mathcal{X} \to \mathbb{R}$ as $h(x) = \min(f(x), g(x))$,

and let $p := d_{\mathrm{TV}}(\mu_1, \mu_2)$. For any $S \in \mathcal{F}$, we let

$$\begin{aligned}
\mu_B(S) &:= \int_{x\in S} \frac{h(x)}{1-p} d\mu_{12}(x), \\
\mu_C(S) &:= \int_{x\in S} \left(h(x)\frac{q-1}{1-p} + f(x) \right) \frac{1}{q} d\mu_{12}(x), \\
\mu_D(S) &:= \int_{x\in S} \left(h(x)\frac{q-1}{1-p} + g(x) \right) \frac{1}{q} d\mu_{12}(x),
\end{aligned}$$

and we set $\chi \sim \text{Bernoulli}(q), B \sim \mu_B, C \sim \mu_C, D \sim \mu_D$ be independent random variables. With this choice, it is straightforward to check that the conditions of the lemma are satisfied. □

5.5.2 Concentration by Stein's method of exchangeable pairs

Let $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a Polish space, and X is a random variable taking values in $\mathcal{X}$. We are interested in the concentration properties of $f(X)$. Suppose that $\mathbb{E}(f(X)) = 0$. Let (X, X') be an exchangeable pair, $m(\theta) := \mathbb{E}(e^{\theta f(X)})$. Suppose that $F(x, y) : \mathcal{X}^2 \to \mathbb{R}$ is an antisymmetric function satisfying

$$\mathbb{E}(F(X, X')|X) = f(X). \tag{5.5.3}$$

Then for any $\theta \in \mathbb{R}$,

$$\begin{aligned}
m'(\theta) &= \mathbb{E}(f(X)e^{\theta f(X)}) = \mathbb{E}(F(X, X')e^{\theta f(X)}) = -\mathbb{E}(F(X, X')e^{\theta f(X')}) \\
&= \mathbb{E}\left(F(X, X')\frac{e^{\theta f(X)} - e^{\theta f(X')}}{2} \right).
\end{aligned} \tag{5.5.4}$$

By Chatterjee (2005), this can be further bounded by

$$\mathbb{E}\left(\frac{\theta}{2}|F(X,X')||f(X)-f(X')|e^{\theta f(X)}\right),$$

and conditions on $\Delta(X) := \frac{1}{2}\mathbb{E}\left(|F(X,X')||f(X)-f(X')|\,X\right)$ determine the concentration properties of $f(X)$.

In this chapter, we are also going to use (5.5.4), but instead of taking absolute value, we consider positive and negative parts.

In order to apply the approach for some function f, we need to find the antisymmetric function $F(x,y)$ such that (5.5.3) is satisfied. Chapter 4 of Chatterjee (2005) finds such an antisymmetric function by a method using a Markov chain, we give a summary below.

An exchangeable pair (X,X') automatically defines a reversible Markov kernel P as

$$Pf(x) := E(f(X')|X=x), \tag{5.5.5}$$

where f is any function such that $\mathbb{E}|f(X)| < \infty$.

Let $\{X(k)\}_{k\geq 0}$ and $\{X'(k)\}_{k\geq 0}$ be two chains with Markov kernel P, having arbitrary initial values, and coupled according to some coupling scheme which satisfies the following property.

P For every initial value (x,y) of the joint chain $\{X(k)\}_{k\geq 0}, \{X'(k)\}_{k\geq 0}$, and every k, the marginal distribution of $X(k)$ depends only on x and the marginal distribution of $X'(k)$ depends only on y.

Under this assumption, the following lemma holds.

Lemma 5.5.2 (Lemma 4.2 of Chatterjee (2005))**.** *Suppose the chains* $\{X(k)\}$ *and* $\{X'(k)\}$ *satisfy the property* $\boldsymbol{P}$ *described above. Let* $f : \mathcal{X} \to \mathbb{R}$ *be a function such that* $\mathbb{E}f(X) = 0$*. Suppose there exists a finite constant* L *such that for every* $(x, y) \in \mathcal{X}^2$*,*

$$\sum_{k=0}^{\infty} |\mathbb{E}(f(X(k)) - f(X'(k))|X(0) = x, X'(0) = y)| \leq L. \tag{5.5.6}$$

Then, the function F*, defined as*

$$F(x, y) := \sum_{k=0}^{\infty} \mathbb{E}(f(X(k)) - f(X'(k))|X(0) = x, X'(0) = y),$$

satisfies $F(X, X') = -F(X', X)$ *and* $\mathbb{E}(F(X, X')|X) = f(X)$*.*

Remark 5.5.3. It is useful to start with $X(0) = X$ and $X'(0) = X'$, because we can bound $F(X, X')$ during the verification of (5.5.6).

5.5.3 Additional lemmas

The following lemma proves concentration in the case when $\Delta(X)$ is not bounded almost surely, but itself is concentrated (a reformulation of Lemma 11 of Massart (2000)). Since the proof is short, we include it for completeness (it is based on part of the proof of Theorem 3.13 of Chatterjee (2005)).

Lemma 5.5.4. *Let* $m(\theta) = \mathbb{E}(e^{\theta f(X)})$*. For any random variable* V*, and any* $L > 0$*, we have for every* $\theta \in \mathbb{R}$*,*

$$\mathbb{E}(e^{\theta f(X)}V) \leq L^{-1} \log \mathbb{E}(e^{LV})m(\theta) + L^{-1}\theta m'(\theta) - L^{-1}m(\theta) \log(m(\theta)),$$

if the expectations on both sides exist.

Proof. Let $u(X) := \frac{e^{\theta f(X)}}{m(\theta)}$. Let $A, B \geq 0$ be two random variables with finite variance and $\mathbb{E}(A) = 1$, then

$$\mathbb{E}(A \log(B)) \leq \log(\mathbb{E}(AB)),$$

which can be shown by changing the measure and applying Jensen's inequality. Using this, we have

$$\begin{aligned} \mathbb{E}(e^{\theta f(X)} V) &= L^{-1} m(\theta) \mathbb{E}\left(u(X) \left(\log \frac{e^{LV}}{u(X)} + \log u(X) \right) \right) \\ &\leq L^{-1} \log \mathbb{E}(e^{LV}) m(\theta) + L^{-1} \mathbb{E}\left(e^{\theta f(X)} \log u(X) \right), \end{aligned}$$

here we applied our previous inequality with $A = u(X)$ and $B = \frac{e^{LV}}{u(X)}$. Now using the fact that $\log(u(X)) = \theta f(X) - \log(m(\theta))$, we obtain the result. □

We will use the following well known result many times in our proofs.

Lemma 5.5.5. *Let W be a centered random variable with moment generating function $m(\theta)$. Let $C, D \geq 0$, suppose that $m(\theta)$ is finite, and continuously differentiable in $[0, 1/C)$, and satisfies*

$$m'(\theta) \leq C\theta m'(\theta) + D\theta m(\theta).$$

Then for $0 \leq \theta < 1/C$,

$$\log(m(\theta)) \leq \frac{D\theta^2}{2(1 - C\theta)}, \tag{5.5.7}$$

and for every $t \geq 0$,

$$\mathbb{P}(W \geq t) \leq \exp\left(-\frac{t^2}{2(D + Ct)} \right). \tag{5.5.8}$$

Proof. By rearranging, we have

$$\begin{aligned}(1-C\theta)m'(\theta) &\leq D\theta m(\theta)\\ \log(m(\theta))' &\leq \frac{D\theta}{1-C\theta}\\ \log(m(\theta)) &\leq \int_{x=0}^{\theta}\frac{Dx}{1-Cx} = -\frac{D\theta}{C} - \frac{D\log(1-C\theta)}{C^2} \leq \frac{D\theta^2}{2(1-C\theta)},\end{aligned}$$

using the fact that for $0 \leq z \leq 1$, $-z - \log(1-z) \leq \frac{z^2}{2(1-z)}$. We obtain the tail bound by applying Markov's inequality for $\theta = \frac{t}{D+Ct}$. □

5.6 Proofs of the main results

In this section, we are going to prove our main result, Theorem 5.3.1 and Corollary 5.3.2. The theorem concerns dependent random variables, and we need to introduce a certain amount of notation to handle them, making the proof rather technical. In order to help the reader in digesting this proof, we are going to prove the theorem first in the independent case, where we are free of the notational burden required for dependent random variables.

Before starting the proof in the independent case, we introduce some notation and two lemmas that are going to be used in both the independent and the dependent cases.

Let $X = (X_1, \ldots, X_n)$ be an vector of random variables taking value in Λ. Let $f : \Lambda \to \mathbb{R}$ be the centered version of g, defined as

$$f(x) = g(x) - \mathbb{E}(g(X)) \text{ for every } x \in \Lambda. \tag{5.6.1}$$

Let $\alpha_1, \ldots, \alpha_n : \Lambda \to \mathbb{R}_+$ be functions such that for any $x, y \in \Lambda$,

$$f(x) - f(y) \leq \sum_{i=1}^{n} \mathbb{1}[x_i \neq y_i]\alpha_i(x); \tag{5.6.2}$$

let $\alpha(x) := (\alpha_1(x), \ldots, \alpha_n(x))$. Note that at this point we do not yet make any specific self-bounding type assumptions on $\alpha(x)$.

Let I be uniformly distributed in $[n]$. Suppose that (X, X') is an exchangeable pair, such that $X_i = X'_i$ for every $i \in [n] \setminus \{I\}$. Suppose that for $k \geq 0$, $X(k)$ and $X'(k)$ are Markov chains with kernel defined as in (5.5.5), satisfying Property **P** and (5.5.6). For $k \geq 0$, define the random vector $L(k) \in \mathbb{R}^n_+$ as

$$L_i(k) := \mathbb{1}[X_i(k) \neq X'_i(k)] \text{ for } 1 \leq i \leq n.$$

The following two lemmas bound the moment generating function of f in function of the vectors $L(k)$ and $\alpha(x)$.

Lemma 5.6.1. *Under the above assumptions, for $\theta > 0$, if $m(\theta) < \infty$, then we have*

$$m'(\theta) \leq \mathbb{E}\left(\sum_{k=0}^{\infty} \langle L(k), \alpha(X(k))\rangle \, \alpha_I(X)\theta e^{\theta f(X)}\right).$$

Proof. Note that

$$\begin{aligned} m'(\theta) &= \mathbb{E}(f(X)e^{\theta f(X)}) \\ &= \mathbb{E}\left(F(X, X')e^{\theta f(X)}\right) = \frac{1}{2}\mathbb{E}\left(F(X, X')(e^{\theta f(X)} - e^{\theta f(X')}\right) \\ &\leq \mathbb{E}\left((F(X, X'))_+(e^{\theta f(X)} - e^{\theta f(X')})_+\right) \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left((F(X,X'))_+(1-e^{-\theta(f(X)-f(X'))_+})e^{\theta f(X)}\right)\\
&\le \mathbb{E}\left((F(X,X'))_+(f(X)-f(X'))_+\theta e^{\theta f(X)}\right)\\
&\le \mathbb{E}\left(\sum_{k=0}^{\infty}\left(f(X(k))-f(X'(k))\right)_+\left(f(X)-f(X')\right)_+\theta e^{\theta f(X)}\right).
\end{aligned}
$$

Using (5.6.2), we have

$$
\left(f(X)-f(X')\right)_+ \le \alpha_I(X), \text{ and } \left(f(X(k))-f(X'(k))\right)_+ \le \langle L(k), \alpha(X(k))\rangle,
$$

thus the result follows. □

Lemma 5.6.2. *Under the above assumptions, for $\theta < 0$, if $m(\theta) < \infty$, and in addition, $f(X) - f(X') \le 1$ almost surely, then*

$$
m'(\theta) \ge -\sum_{k=0}^{\infty}\mathbb{E}\left(\left(e^{-\theta}-1\right)e^{\theta f(X)}\langle L(k),\alpha(X(k))\rangle\,\alpha_I\right).
$$

Proof. Note that

$$
\begin{aligned}
m'(\theta) &= \frac{1}{2}\mathbb{E}\left(F(X,X')\left(e^{\theta f(X)}-e^{\theta f(X')}\right)\right)\\
&\ge -\mathbb{E}\left((F(X,X'))_+\left(e^{\theta f(X)}-e^{\theta f(X')}\right)_-\right)\\
&\ge -\mathbb{E}\left((F(X,X'))_+\left(e^{\theta f(X')}-e^{\theta f(X)}\right)_+\right)\\
&\ge -\mathbb{E}\left((F(X,X'))_+\left(e^{\theta(f(X')-f(X))}-1\right)_+e^{\theta f(X)}\right)\\
&= -\mathbb{E}\left((F(X,X'))_+\left(e^{-\theta(f(X)-f(X'))_+}-1\right)e^{\theta f(X)}\right).
\end{aligned}
$$

Since $\theta < 0$, and $\left(e^{(-\theta)x}-1\right)/x$ is a monotone function in x for $x \ge 0$, using $0 \le$

$(f(X) - f(X'))_+ \leq 1$, we obtain

$$\left(e^{-\theta(f(X)-f(X'))_+} - 1\right) \leq (f(X) - f(X'))_+ \left(e^{-\theta} - 1\right).$$

Now applying (5.6.2) proves the result. □

5.6.1 Independent case

In this section, we are going to prove Theorem 5.3.1 and Corollary 5.3.2 under the additional assumption that $X = (X_1, \ldots, X_n)$ is a vector independent random variables. First, we are going to construct a valid coupling of $(X(k), X'(k))_{k\geq 0}$, satisfying Property **P** and (5.5.6). After this, we will use Lemma 5.6.1 and 5.6.1 to obtain the mgf bounds of Theorem 5.3.1.

The construction of $(X(k), X'(k))_{k\geq 0}$ is the same as in Example on page 73 of Chatterjee (2005), sketched here for the sake of completeness. This is a version of the Glauber dynamics. First, we set $X(0) = x$, and $X'(0) = y$ for some $x, y \in \Lambda$. Then we let $I(1), I(2), \ldots$ be independent random variables uniformly distributed on $[n]$, and $X^*(1), X^*(2), \ldots$ be independent copies of X. Then in the first step, we define the vectors $X(1)$ and $X'(1)$ as equal to $X(0)$, and $X'(0)$, respectively, except in coordinate $I(1)$, where we set $X_{I(1)}(1) = X'_{I(1)}(1) = X^*_{I(1)}(1)$. We define $X(k), X'(k)$ in the same way, by starting from $X(k-1), X'(k-1)$, and changing their coordinate $I(k)$ to $X^*_{I(k)}(k)$. This coupling has shown to satisfy Property **P** and (5.5.6) in Chatterjee (2005) (via the coupon collector's problem). Finally, we note that X' is defined as one step in the dynamics, that is, we let X^* be an independent copy of X, I be uniformly distributed on $[n]$, independently of X and X^*, and X' equals to X except in coordinate I, where it equals X^*_I.

Now we are ready to prove Theorem 5.3.1 and Corollary 5.3.2 under the independence assumption.

Proof of Part 1 of Theorem 5.3.1 and Corollary 5.3.2 assuming independence.
By Lemma 5.6.1, using the fact that f is bounded under our assumptions, we have that for $\theta > 0$,

$$m'(\theta) \leq \sum_{k=0}^{\infty} \mathbb{E}\left(\theta e^{\theta f(X)} \cdot \sum_{i=1}^{n} \alpha_i(X(k))\alpha_i(X)\mathbb{1}[i \notin I(1), \ldots, I(k)]\right)$$

Now by our assumption, $\alpha_i(X(k)) \leq 1$, and using that g is (a,b)-*-self-bounding,

$$\begin{aligned}
m'(\theta) &\leq \sum_{k=0}^{\infty} \mathbb{E}\left(\theta e^{\theta f(X)} \cdot \frac{1}{n}\sum_{i=1}^{n} \alpha_i(X)\mathbb{1}[i \notin I(1), \ldots, I(k)]\right) \\
&\leq \mathbb{E}\left(\theta e^{\theta f(X)} \cdot \frac{1}{n}\sum_{i=1}^{n} \alpha_i(X) \sum_{k=0}^{\infty}\left(1 - \frac{1}{n}\right)^k\right) \\
&\leq \mathbb{E}\left(\theta e^{\theta f(X)}(ag(X) + b)\right) = \mathbb{E}\left(\theta e^{\theta f(X)}(af(X) + (a\mathbb{E}g(X) + b))\right) \\
&\leq \theta a m'(\theta) + \theta\left(a\mathbb{E}g(X) + b\right) m(\theta).
\end{aligned}$$

The mgf bound now follows by rearrangement and integration, and applying Lemma 5.5.5 proves the concentration bound of Corollary 5.3.2. □

Proof of Part 2 of Theorem 5.3.1 and Corollary 5.3.2 assuming independence.
By Lemma 5.6.1, we have for $\theta > 0$

$$m'(\theta) \leq \sum_{k=0}^{\infty} \mathbb{E}\left(\theta e^{\theta f(X)} \cdot \frac{1}{n}\sum_{i=1}^{n} \alpha_i(X(k))\alpha_i(X)\mathbb{1}[i \notin I(1), \ldots, I(k)]\right). \tag{5.6.3}$$

Now by the fact that g is weakly (a, b)-*-self-bounding, we have

$$\sum_{i=1}^{n} \alpha_i(X)^2 \leq ag(X) + b, \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i(X(k))^2 \leq ag(X(k)) + b.$$

We will use the conditional version of the Cauchy-Schwarz inequality: if A_i, B_i are random variables for $1 \leq i \leq n$, then

$$\mathbb{E}(A_i B_i|X) \leq \left(\mathbb{E}(A_i^2|X)\right)^{1/2} \cdot \left(\mathbb{E}(B_i^2|X)\right)^{1/2},$$
$$\mathbb{E}\left(\sum_{i=1}^{n} A_i B_i \middle| X\right) \leq \sum_{i=1}^{n} \left(\mathbb{E}(A_i^2|X)\right)^{1/2} \cdot \left(\mathbb{E}(B_i^2|X)\right)^{1/2}.$$

Now writing $A_i = \alpha_i(X)\mathbb{1}[i \notin I(1), \ldots, I(k)]$ and $B_i = \alpha_i(X(k))$, we obtain

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}(\alpha_i(X(k))\alpha_i(X)\mathbb{1}[i \notin I(1), \ldots, I(k)]|X) \\
&\leq \sum_{i=1}^{n} \left(\mathbb{E}(\alpha_i(X)^2\mathbb{1}[i \notin I(1), \ldots, I(k)]|X)\right)^{1/2} \cdot \left(\mathbb{E}(\alpha_i(X(k))^2|X)\right)^{1/2} \\
&= \left(1 - \frac{1}{n}\right)^{k/2} \cdot \sum_{i=1}^{n} (\alpha_i(X)^2)^{1/2} \cdot \left(\mathbb{E}(\alpha_i(X(k))^2|X)\right)^{1/2} \\
&\leq \left(1 - \frac{1}{n}\right)^{k/2} \cdot \sum_{i=1}^{n} \frac{1}{2}\mathbb{E}\left(\alpha_i(X)^2 + \alpha_i(X(k))^2|X\right) \\
&\leq \left(1 - \frac{1}{n}\right)^{k/2} \cdot \frac{1}{2}\mathbb{E}(ag(X) + b + ag(X(k)) + b|X)
\end{aligned}$$

By substituting this into (5.6.3), we obtain that

$$m'(\theta) \leq \sum_{k=0}^{\infty} \mathbb{E}\left(\theta e^{\theta f(X)} \frac{1}{n} \sum_{k=0}^{\infty} \left(1 - \frac{1}{n}\right)^{k/2} \frac{1}{2}(ag(X) + b + ag(X(k)) + b)\right)$$

$$\leq \sum_{k=0}^{\infty} \mathbb{E}\left(\theta e^{\theta f(X)} \frac{1}{n} \sum_{k=0}^{\infty} \left(1-\frac{1}{n}\right)^{k/2} (ag(X)+b)\right)$$
$$\leq \mathbb{E}\left(\theta e^{\theta f(X)} 2(ag(X)+b)\right) = \mathbb{E}\left(\theta e^{\theta f(X)}(2af(X)+2a\mathbb{E}g(X)+2b)\right)$$
$$\leq \theta 2am'(\theta) + \theta\left(2a\mathbb{E}g(X)+2b\right) m(\theta).$$

Here we have used the fact that for $\theta > 0$,

$$\mathbb{E}(e^{\theta f(X)} f(X(k))) \leq \mathbb{E}(e^{\theta f(X)} f(X)), \tag{5.6.4}$$

since using the exchangeability of $f(X)$ and $f(X(k))$,

$$\mathbb{E}\left(e^{\theta f(X)}\left(f(X)-f(X(k))\right)\right) = \mathbb{E}\left(e^{\theta f(X(k))}(f(X(k))-f(X))\right)$$
$$= \mathbb{E}\left(\left(e^{\theta f(X)} - e^{\theta f(X(k))}\right)\left(f(X)-f(X(k))\right)\right) \geq 0,$$

since $e^{\theta f(X)} - e^{\theta f(X(k))}$ and $f(X) - f(X(k))$ always have the same sign. We conclude by applying Lemma 5.5.5. □

Proof of Part 3 of Theorem 5.3.1 and Corollary 5.3.2 assuming independence.
By Lemma 5.6.2,

$$m'(\theta) \geq -\sum_{k=0}^{\infty} \mathbb{E}\left(\left(e^{-\theta}-1\right) e^{\theta f(X)} \cdot \frac{1}{n} \sum_{i=1}^{n} \alpha_i(X(k))\alpha_i(X)\mathbb{1}[i \notin I(1), \ldots, I(k)]\right).$$

In Part 2, we proved that

$$\sum_{i=1}^{n} \mathbb{E}(\alpha_i(X(k))\alpha_i(X)\mathbb{1}[i \notin I(1), \ldots, I(k)]|X)$$
$$\leq \left(1 - \frac{1}{n}\right)^{k/2} \cdot \frac{1}{2}\mathbb{E}(ag(X) + b + ag(X(k)) + b|X),$$

so we obtain

$$m'(\theta) \geq -\mathbb{E}\left(\left(e^{-\theta} - 1\right) e^{\theta f(X)} \frac{1}{n}\right. \tag{5.6.5}$$
$$\left. \cdot \sum_{k=0}^{\infty} \left(1 - \frac{1}{n}\right)^{k/2} \cdot \frac{1}{2}\left(af(X) + af(X(k)) + 2b + 2a\mathbb{E}g(X)\right)\right).$$

The terms involving $f(X(k))$ cause some difficulty. Although we can show, in the same way as in Part 2, that

$$-\mathbb{E}(e^{\theta f(X)} f(X(k))) \leq -\mathbb{E}(e^{\theta f(X)} f(X)),$$

for us the other sided inequality would be more convenient. Nevertheless, we can use the concentration properties of $f(X(k))$ from Part 2 to bound this term. By Lemma 5.5.4, for any $L > 0$,

$$\mathbb{E}(e^{\theta f(X)} f(X(k))) \leq L^{-1} \log \mathbb{E}(e^{Lf(X(k))}) m(\theta) + L^{-1}\theta m'(\theta)$$

Now by exchangeability $\mathbb{E}(e^{Lf(X(k))}) = \mathbb{E}(e^{Lf(X)}) = m(L)$, and we can use the bound

from Part 2 to obtain that for $0 < L < 1/(2a)$,

$$\begin{aligned}
&\log(m(L)) \leq \frac{(a\mathbb{E}g(X)+b)L^2}{(1-2aL)} \\
&\mathbb{E}(e^{\theta f(X)} f(X(k))) \leq \frac{(a\mathbb{E}g(X)+b)L}{(1-2aL)} m(\theta) + L^{-1}\theta m'(\theta) \\
&= \mathbb{E}\left[\frac{(a\mathbb{E}g(X)+b)L}{(1-2aL)} e^{\theta f(X)} + L^{-1}\theta f(X) e^{\theta f(X)}\right]
\end{aligned}$$

Substituting this back to (5.6.5), and summing up in k as previously, we obtain

$$\begin{aligned}
&m'(\theta) \geq -\left(e^{-\theta}-1\right) \\
&\cdot \mathbb{E}\left[e^{\theta f(X)}\left(2a\mathbb{E}g(X)+2b+a\frac{(a\mathbb{E}g(X)+b)L}{(1-2aL)}\right) + f(X)e^{\theta f(X)}\left(a+aL^{-1}\theta\right)\right].
\end{aligned}$$

A convenient choice for L, which makes the inequality tractable, is $L = -\theta$. With this choice, for $0 > \theta > -\frac{1}{2a}$, we obtain

$$\begin{aligned}
m'(\theta) &\geq -\left(e^{-\theta}-1\right)\left(2a\mathbb{E}g(X)+2b-a\frac{(a\mathbb{E}g(X)+b)\theta}{(1+2a\theta)}\right) m(\theta) \\
\log(m(\theta))' &\geq -\left(e^{-\theta}-1\right)\left(2a\mathbb{E}g(X)+2b-a\frac{(a\mathbb{E}g(X)+b)\theta}{(1+2a\theta)}\right),
\end{aligned}$$

thus we have shown (5.3.2). Now we turn to the proof of the concentration bounds of Corollary 5.3.2. Suppose that $0 > \theta > -\frac{1}{4a}$, then $1 + 2a\theta \geq 1/2$, so

$$\log(m(\theta))' \geq -\left(e^{-\theta}-1\right)(2-2a\theta)(a\mathbb{E}g(X)+b). \tag{5.6.6}$$

Now we consider two cases, depending on the size of a. The function $\left(e^x-1\right)/x$ is

increasing for positive x, so we can write

$$
\begin{aligned}
-\left(e^{-\theta}-1\right)(2-2a\theta) &\geq \frac{\left(e^{\frac{1}{4a}}-1\right)}{1/(4a)}\frac{5}{2}\theta \\
\log(m(\theta))' &\geq \frac{\left(e^{\frac{1}{4a}}-1\right)}{1/(4a)}\frac{5}{2}\theta(a\mathbb{E}g(X)+b) \\
\log(m(\theta)) &\leq \frac{\left(e^{\frac{1}{4a}}-1\right)}{1/(4a)}\frac{5}{4}(a\mathbb{E}g(X)+b)\theta^2 \leq 2(a\mathbb{E}g(X)+b)\theta^2,
\end{aligned}
$$

whenever

$$
\frac{\left(e^{\frac{1}{4a}}-1\right)}{1/(4a)} \leq \frac{8}{5}, \tag{5.6.7}
$$

that is, whenever $a \geq a_c$ (with a_c defined as in (5.3.3)). Using Markov's inequality, we have that for $0 < t < \mathbb{E}g(X)$, $0 > \theta > -\frac{1}{4a}$,

$$
\log \mathbb{P}(f(X) \leq -t) \leq \log(m(\theta)) + t\theta \leq 2(a\mathbb{E}g(X)+b)\theta^2 + \theta t,
$$

which takes its minimum at

$$
\theta_{min} = \frac{-t}{4(a\mathbb{E}g(X)+b)},
$$

which satisfies $0 > \theta > -\frac{1}{4a}$, and thus

$$
\log \mathbb{P}(f(X) \leq -t) \leq \frac{-t^2}{8(a\mathbb{E}g(X)+b)}.
$$

Finally, we need to tackle the case when $a < a_c$. Going back to equation (5.6.6), we

can write that for $0 > \theta > -\frac{1}{4a}$,

$$\log(m(\theta))' \geq -\left(e^{-\theta} - 1\right)\frac{5}{2}(a\mathbb{E}g(X) + b)$$
$$\log(m(\theta)) \leq \left(e^{-\theta} + \theta - 1\right)\frac{5}{2}(a\mathbb{E}g(X) + b)$$

Let us write $C := \frac{5}{2}(a\mathbb{E}g(X) + b)$, then by Markov's inequality, we have that for $0 > \theta > -\frac{1}{4a}$, $0 < t < \mathbb{E}g(X)$,

$$\log(\mathbb{P}(f(X) \leq -t)) \leq \log(m(\theta)) + \theta t \leq \left(e^{-\theta} + \theta - 1\right)C + \theta t$$

The minimum of the right hand side is taken at

$$\theta_{min} = -\log\left(1 + \frac{t}{C}\right) \geq -\log\left(1 + \frac{2}{5}\cdot\frac{1}{a}\right),$$

which satisfies $0 > \theta_{min} > -\frac{1}{4a}$ whenever $a < a_c$. Thus, in this case we have

$$\log(\mathbb{P}(f(X) \leq -t)) \leq \left(\frac{t}{C} - \log\left(1 + \frac{t}{C}\right)\right)C - \log\left(1 + \frac{t}{C}\right)t$$
$$= C\left[\frac{t}{C} - \log\left(1 + \frac{t}{C}\right)\left(1 + \frac{t}{C}\right)\right].$$

Now let us take a look at the $x - \log(1 + x)(1 + x)$ function for positive x, we can easily check that this is negative, and

$$x - \log(1 + x)(1 + x) \leq -\frac{x^2}{2 + (2/3)x},$$

so

$$\log(\mathbb{P}(f(X) \leq -t)) \leq -\frac{t^2}{2C + (2/3)t} = -\frac{t^2}{5(a\mathbb{E}g(X) + b) + (2/3)t}. \qquad \square$$

Discussion

When compared to the original proof of Theorem 4.3 of Chatterjee (2005), we have introduced several new ideas in the proof. Firstly, instead of bounding

$$\Delta(X) := \frac{1}{2}\mathbb{E}(|F(X, X')(f(X) - f(X'))||X),$$

we use the one sided version $(F(X, X'))_+(f(X) - f(X'))_+$. Moreover, we have not taken the expectation of this quantity with respect to X, but instead used a tricky symmetrisation argument in (5.6.12). Finally, we have also used Lemma 5.5.4, which was not needed for the original proof. In an upcoming paper, we are going to show that these techniques are powerful enough to imply the exponential and polynomial Efron-Stein inequalities for independent random variables, due to Boucheron, Lugosi, and Massart (2003) and Boucheron, Bousquet, Lugosi, and Massart (2005b). The dependent case remains an open problem.

5.6.2 Dependent case

In this section, we are going to prove Theorem 5.3.1 and Corollary 5.3.2. First, we will clarify the notations in this section. After this, we state two basic lemmas, and a coupling scheme that will be used in the proof. Finally, we give the proof of the results.

Let $X = (X_1, \ldots, X_n)$ be an vector of random variables taking value in Λ, with Dobrushin interdependence matrix $A = (a_{i,j})_{1 \leq i,j \leq n}$.

Now we will construct a coupling for $\{X(k)\}_{k\geq 0}$, and $\{X'(k)\}_{k\geq 0}$. Suppose that we have already coupled

$$X(0),\dots,X(k) \quad \text{and} \quad X'(0),\dots,X'(k),$$

and that $X(k)=x$, $X'(k)=y$. Let $I(k+1)$ be uniformly chosen from $[n]$, independently of the previously defined variables. In order to obtain $X_{I(k+1)}(k+1)$ and $X'_{I(k+1)}(k+1)$, write

$$\nu_1 := \mu_{I(k+1)}(\cdot|x_{-I(k+1)}) \quad \text{and} \quad \nu_2 := \mu_{I(k+1)}(\cdot|y_{-I(k+1)}).$$

By Lemma 5.5.1, we can define the same way as in Section 5.5.1, there exists $B(k+1)$, $C(k+1)$, $D(k+1)$, $\chi(k+1)$ conditionally independent of each other given $X_{-I(k+1)}(k)$ and $X'_{-I(k+1)}(k)$. We can choose $\chi(k+1)\sim$ Bernoulli(q) for any $q\geq d_{\mathrm{TV}}(\nu_1,\nu_2)$.

Let $\xi(k+1)$ be a random vector taking values in $\{0,1\}^n$, having distribution

$$\xi(k+1) := e_i \text{ with probability } a_{I(k+1),i} \ (i\in[n]), \text{ otherwise } \xi(k+1):=0, \quad (5.6.8)$$

where $e_i=(0,\dots,0,1,0,\dots,0)$ is the ith unit vector, and by 0 we mean the null vector. We suppose that $\xi(k+1)$ is conditionally independent of all else given $I(k+1)$. This distribution exists, since

$$\sum_{i=1}^{n} a_{I(k+1),i} \leq \|A\|_\infty \leq 1,$$

by our assumptions. Define

$$\chi(k+1) := \langle \xi(k+1), L(k) \rangle \,, \tag{5.6.9}$$

with $\langle \cdot, \cdot \rangle$ denoting scalar product. Then $\chi(k+1) \sim \text{Bernoulli}(q)$ with

$$q = \sum_{i=1}^{n} a_{I(k+1),i} L_i(k) \geq d_{\text{TV}}(\nu_1, \nu_2).$$

Note that we may have $q > d_{\text{TV}}(\nu_1, \nu_2)$, thus our coupling is different from "the greedy coupling" that is used on page 76 of Chatterjee (2005).

By Lemma 5.5.1, we can define

$$X_{I(k+1)}(k+1) := (1 - \chi(k+1))B(k+1) + \chi(k+1)C(k+1),$$

and

$$X'_{I(k+1)}(k+1) := (1 - \chi(k+1))B(k+1) + \chi(k+1)D(k+1),$$

for all $i \neq I(k+1)$, $X_i(k+1) := X_i(k)$ and $X'_i(k+1) := X'_i(k)$. It is easy to verify by induction that this coupling scheme satisfies Property **P**. For a vector $v \in \mathbb{R}^n$, and $i \in [n]$, define $M(i, v)$ as an $n \times n$ matrix, with $(M(i, v))_{l,m} = \mathbb{1}[l = m]$ for every $1 \leq l, m \leq n$ such that $l \neq i$, and $(M(i, v))_{i,m} = v_m$ for every $1 \leq m \leq n$ (thus it equals to the identity matrix in every row except the ith one where it equals to v).

For example,

$$M(3,(1,0,0,0,0)) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The following lemma states a recursive bound for $L(k)$.

Lemma 5.6.3. *For the above coupling, for every $k \geq 0$*

$$L(k+1) \leq M(I(k+1), \xi(k+1))L(k), \tag{5.6.10}$$

and thus

$$L(k) \leq M(I(k), \xi(k)) \dots M(I(1), \xi(1))L(0). \tag{5.6.11}$$

Proof. Because of the construction of the coupling, we have $L_i(k) = L_i(k+1)$ if $i \neq I(k+1)$. Moreover, $X_{I(k+1)}(k+1) \neq X'_{I(k+1)}(k+1)$ implies that $\chi(k+1) = 1$, so (5.6.10) follows by the definitions of $\chi(k+1)$ and $M(I(k+1), \xi(k+1))$. We obtain (5.6.11) by iteration. □

Note that in Theorem 5.3.1, in each of the three cases, g is always going to be bounded, thus f is also bounded. This means that we have $|f(x)| \leq C$ for some absolute constant C for every $x \in \Lambda$. Using this and (5.6.11), we have

$$\begin{aligned} &|\mathbb{E}(f(X(k)) - f(X'(k))|X(0) = x, X'(0) = y)| \\ &\quad \leq \mathbb{E}(2C\|L(k)\|_1 | X(0) = x, X'(0) = y) \leq 2C\|[\mathbb{E}(M(I(1), \xi(1)))]^k\|_1 \|L(0)\|_1 \end{aligned}$$

$$\leq 2nC \left\| \left(1 - \frac{1}{n}E + \frac{1}{n}A\right)^k \right\|_1 \leq 2nC \left(1 - \frac{1}{n} + \frac{1}{n}\|A\|_1\right)^k,$$

so by summing up, we obtain that (5.5.6) holds with $L = 2nC/\left(1 - \frac{1}{n} + \frac{1}{n}\|A\|_1\right)$. Now we are ready to prove Theorem 5.3.1 and Corollary 5.3.2.

Proof of Part 1 of Theorem 5.3.1 and Corollary 5.3.2. For $\theta > 0$, using Lemma 5.6.1, we have

$$m'(\theta) \leq \mathbb{E}\left(\sum_{k=0}^{\infty} \langle L(k), \alpha(X(k))\rangle \, \alpha_I(X) \theta e^{\theta f(X)}\right).$$

Let $\{X(k), X'(k)\}_{k\geq 0}$ be defined as in our coupling scheme, then using (5.6.11), and the fact that $L(0) \leq e_I$, we can write

$$\begin{aligned}
&\mathbb{E}\left(\langle L(k), \alpha(X(k))\rangle \, \alpha_I(X) \middle| X\right) \\
&\quad \leq \mathbb{E}\left(\langle (M(I(k), \xi(k)) \dots M(I(1), \xi(1)) e_I), \alpha(X(k))\rangle \, \alpha_I(X) \middle| X\right) \\
&\quad \leq \frac{1}{n}\mathbb{E}\left(\alpha(X(k))^t \left(M(I(k), \xi(k)) \dots M(I(1), \xi(1))\right) \alpha(X) \middle| X\right) \\
&\quad \leq \frac{1}{n}\mathbb{E}\left(\|\alpha(X(k))\|_\infty \|M(I(k), \xi(k)) \dots M(I(1), \xi(1)) \alpha(X)\|_1 \middle| X\right).
\end{aligned}$$

Denote by E the identity matrix of size n. Using the facts that for *-self-bounding functions, $\|\alpha(X(k))\|_\infty \leq 1$, and that the elements of $M(I(k), \xi(k))$ and $L(k)$ are non-negative for every k, we obtain

$$\begin{aligned}
&\mathbb{E}\left(\langle L(k), \alpha(X(k))\rangle \, \alpha_I(X) \middle| X\right) \\
&\quad \leq \mathbb{E}\left(\langle M(I(k), \xi(k)) \dots M(I(1), \xi(1)) e_I, 1\rangle \, \alpha_I(X) \middle| X\right),
\end{aligned}$$

with 1 denoting an n vector of ones. Using the fact that $M(I(1), \xi(1))$, …, $M(I(k), \xi(k))$

are independent of I and X, we have

$$\begin{aligned}
&\mathbb{E}\left(\left\langle L(k), \alpha(X(k))\right\rangle \alpha_I(X) \middle| X\right) \\
&\quad \leq \frac{1}{n} \left\|\mathbb{E}\left(M(I(k), \xi(k)) \dots M(I(1), \xi(1)) \middle| X\right)\right\|_1 \|\alpha(X)\|_1 \\
&\quad \leq \frac{1}{n} \left\|\mathbb{E}\left(M(I(1), \xi(1)) \middle| X\right)^k\right\|_1 (ag(X) + b) \\
&\quad \leq \frac{1}{n} \left\|\left(\left(1 - \frac{1}{n}\right) E + \frac{1}{n} A\right)^k\right\|_1 (ag(X) + b) \\
&\quad \leq \frac{1}{n} \left(1 - \frac{1}{n} + \frac{1}{n}\|A\|_1\right)^k (af(X) + a\mathbb{E}(g) + b),
\end{aligned}$$

We sum up in k, and obtain that

$$\begin{aligned}
m'(\theta) &\leq \sum_{k=0}^{\infty} \frac{1}{n} \left(1 - \frac{1}{n} + \frac{1}{n}\|A\|_1\right)^k \mathbb{E}\left((af(X) + a\mathbb{E}(g) + b)\theta e^{\theta f(X)}\right), \\
m'(\theta) &\leq \frac{1}{1 - \|A\|_1} \left(a\theta m'(\theta) + (a\mathbb{E}(g) + b)\theta m(\theta)\right).
\end{aligned}$$

We obtain the mgf bound in Theorem 5.3.1 by integration of this inequality, and our concentration bound in Corollary 5.3.2 from Lemma 5.5.5. □

Proof of Part 2 of Theorem 5.3.1 and Corollary 5.3.2. As in Part 1, we have that for $\theta > 0$, $m'(\theta) \leq \mathbb{E}\left(\sum_{k=0}^{\infty} \left\langle L(k), \alpha(X(k))\right\rangle \alpha_I(X) \theta e^{\theta f(X)}\right)$, and

$$\begin{aligned}
&\mathbb{E}\left(\left\langle L(k), \alpha(X(k))\right\rangle \alpha_I(X) \middle| X\right) \\
&\leq \frac{1}{n} \mathbb{E}\left(\alpha(X(k))^t \left(M(I(k), \xi(k)) \dots M(I(1), \xi(1))\right) \alpha(X) \middle| X\right) \\
&\leq \frac{1}{n} \mathbb{E}\left(\|\alpha(X(k))\|_2 \left\|M(I(k), \xi(k)) \dots M(I(1), \xi(1)) \alpha(X)\right\|_2 \middle| X\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{n}\mathbb{E}\left(\|\alpha(X(k))\|_2^2\middle| X\right)^{1/2}\mathbb{E}\left(\|M(I(k),\xi(k))\dots M(I(1),\xi(1))\alpha(X)\|_2^2\middle| X\right)^{1/2}\\
&\leq \frac{1}{n}\mathbb{E}\left(ag(X(k))+b\middle| X\right)^{1/2}\cdot\mathbb{E}\Bigg(\alpha(X)^tM(I(1),\xi(1))^t\cdot\ldots\cdot M(I(k),\xi(k))^t\\
&\times M(I(k),\xi(k))\cdot\ldots\cdot M(I(1),\xi(1))\alpha(X)\Bigg|X\Bigg)^{1/2}\\
&\leq \frac{1}{n}\mathbb{E}\left(ag(X(k))+b\middle| X\right)^{1/2}\cdot\Bigg(\alpha(X)^t\mathbb{E}\Bigg(M(I(1),\xi(1))^t\cdot\ldots\cdot M(I(k),\xi(k))^t\\
&\times M(I(k),\xi(k))\cdot\ldots\cdot M(I(1),\xi(1))\Bigg|X\Bigg)\alpha(X)\Bigg)^{1/2}\\
&\leq \frac{1}{n}\mathbb{E}\left(ag(X(k))+b\middle| X\right)^{1/2}(ag(X)+b)^{1/2}\\
&\times\left\|\mathbb{E}\big(M(I(1),\xi(1))^t\cdot\ldots\cdot M(I(k),\xi(k))^tM(I(k),\xi(k))\cdot\ldots\cdot M(I(1),\xi(1))\big|X\big)\right\|_2^{1/2}.
\end{aligned}
$$

Now for example

$$
\begin{aligned}
&M(3,(1,0,0,0,0))^t\cdot M(3,(1,0,0,0,0))\\
&=\begin{pmatrix}1&0&1&0&0\\0&1&0&0&0\\0&0&0&0&0\\0&0&0&1&0\\0&0&0&0&1\end{pmatrix}\cdot\begin{pmatrix}1&0&0&0&0\\0&1&0&0&0\\1&0&0&0&0\\0&0&0&1&0\\0&0&0&0&1\end{pmatrix}=\begin{pmatrix}2&0&0&0&0\\0&1&0&0&0\\0&0&0&0&0\\0&0&0&1&0\\0&0&0&0&1\end{pmatrix},
\end{aligned}
$$

so $M(I(k),\xi(k))^tM(I(k),\xi(k))$ is diagonal, therefore it is easy to see that

$$M(I(1),\xi(1))^t\dots M(I(k),\xi(k))^tM(I(k),\xi(k))\dots M(I(1),\xi(1))$$

is also diagonal. Moreover, by denoting the $n\times n$ matrix of only one 1 at position

i, j and zeros elsewhere by $H(i,j)$ and $H(i) := H(i,i)$, we can write

$$\begin{aligned}
&\mathbb{E}(M(I(k),\xi(k))^t M(I(k),\xi(k))|X, I(1),\xi(1),\ldots,I(k-1),\xi(k-1)) \\
&= \mathbb{E}(M(I(k),\xi(k))^t M(I(k),\xi(k))|X) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{n} a_{i,j}(E - H(i) + H(i,j))^t(E - H(i) + H(i,j))\right. \\
&+ \left.\left(1 - \sum_{j=1}^{n} a_{i,j}\right)(E - H(i))^t(E - H(i))\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{n} a_{i,j}(E - H(i) + H(j)) + \left(1 - \sum_{j=1}^{n} a_{i,j}\right)(E - H(i))\right] \\
&= \left(1 - \frac{1}{n}\right)E + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}H(j) = \left(1 - \frac{1}{n}\right)E + \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{n} a_{i,j}\right)H(j).
\end{aligned}$$

Now using the conditions of our theorem, we have $(\sum_{i=1}^{n} a_{i,j}) \leq \|A\|_1 < 1$, so we can write

$$\begin{aligned}
&\mathbb{E}(M(I(k),\xi(k))^t M(I(k),\xi(k))|X, I(1),\xi(1),\ldots,I(k-1),\xi(k-1)) \\
&\leq \left(1 - \frac{1}{n} + \frac{1}{n}\|A\|_1\right)E.
\end{aligned}$$

By repeating this, we obtain that

$$\begin{aligned}
&\left\|\mathbb{E}\left(M(I(1),\xi(1))^t \cdot \ldots \cdot M(I(k),\xi(k))^t M(I(k),\xi(k)) \cdot \ldots \cdot M(I(1),\xi(1))\middle| X\right)\right\|_2^{1/2} \\
&\leq \left(1 - \frac{1}{n} + \frac{1}{n}\|A\|_1\right)^{k/2},
\end{aligned}$$

so summing up in k, we have

$$
\begin{aligned}
& m'(\theta) \\
& \leq \frac{1}{n}\mathbb{E}\left(\sum_{k=0}^{\infty}\mathbb{E}\left(ag(X(k))+b|\, X\right)^{1/2}\left(ag(X)+b\right)^{1/2}\cdot\left(1-\frac{1}{n}+\frac{1}{n}\|A\|_1\right)^{k/2}\theta e^{\theta f(X)}\right) \\
& \leq \frac{1}{n}\mathbb{E}\left(\sum_{k=0}^{\infty}\left(\frac{af(X(k))+af(X)+2b+2a\mathbb{E}(g)}{2}\right)\cdot\left(1-\frac{1}{n}+\frac{1}{n}\|A\|_1\right)^{k/2}\theta e^{\theta f(X)}\right) \\
& \leq \frac{1}{n}\mathbb{E}\left(\sum_{k=0}^{\infty}\left(af(X)+b+a\mathbb{E}(g)\right)\left(1-\frac{1}{n}+\frac{1}{n}\|A\|_1\right)^{k/2}\theta e^{\theta f(X)}\right) \\
& \leq \mathbb{E}\left(\frac{2}{1-\|A\|_1}\left(af(X)+b+a\mathbb{E}(g)\right)\theta e^{\theta f(X)}\right),
\end{aligned}
$$

and the mgf bound in Theorem 5.3.1 follows by integration. Here we have used the fact that for $\theta > 0$,

$$\mathbb{E}(e^{\theta f(X)}f(X(k))) \leq \mathbb{E}(e^{\theta f(X)}f(X)), \tag{5.6.12}$$

because using the exchangeability of $f(X)$ and $f(X(k))$,

$$
\begin{aligned}
& \mathbb{E}\left(e^{\theta f(X)}\left(f(X)-f(X(k))\right)\right) = \mathbb{E}\left(e^{\theta f(X(k))}(f(X(k))-f(X))\right) \\
& = \frac{1}{2}\mathbb{E}\left(\left(e^{\theta f(X)}-e^{\theta f(X(k))}\right)\left(f(X)-f(X(k))\right)\right) \geq 0,
\end{aligned}
$$

since $e^{\theta f(X)}-e^{\theta f(X(k))}$ and $f(X)-f(X(k))$ always have the same sign. Applying Lemma 5.5.5 with $C=\frac{2a}{1-\|A\|_1}$ and $D=\frac{2(a\mathbb{E}(g)+b)}{1-\|A\|_1}$ proves tail inequality in Corollary 5.3.2. □

Proof of Part 3 of Theorem 5.3.1 and Corollary 5.3.2. Now we will bound the lower

tail, so suppose that $\theta < 0$. By Lemma 5.6.2,

$$m'(\theta) \geq -\sum_{k=0}^{\infty} \mathbb{E}\left(\left(e^{-\theta}-1\right) e^{\theta f(X)} \left\langle L(k), \alpha(X(k))\right\rangle \alpha_I\right).$$

In Part 2, we proved that

$$\begin{aligned}
&\mathbb{E}\left(\left\langle L(k), \alpha(X(k))\right\rangle \alpha_I(X) \middle| X\right) \\
&\leq \frac{1}{n}\mathbb{E}\left(\frac{af(X(k)) + af(X) + 2b + 2a\mathbb{E}(g)}{2}\middle| X\right)\left(1 - \frac{1}{n} + \frac{1}{n}\|A\|_1\right)^{k/2}.
\end{aligned}$$

By summing up in k, we obtain

$$\begin{aligned}
m'(\theta) \geq &-\left(e^{-\theta}-1\right)\sum_{k=0}^{\infty}\frac{1}{n}\left(1 - \frac{1}{n} + \frac{1}{n}\|A\|_1\right)^{k/2} \\
&\times \mathbb{E}\left(\left(\frac{af(X(k)) + af(X) + 2b + 2a\mathbb{E}(g)}{2}\right)e^{\theta f(X)}\right).
\end{aligned}$$

By Lemma 5.5.4, since $m(\theta) \geq 1$, for any $L > 0$,

$$\mathbb{E}(e^{\theta f(X)} f(X(k))) \leq L^{-1}\log \mathbb{E}(e^{Lf(X(k))})m(\theta) + L^{-1}\theta m'(\theta),$$

and by Part 2, for $0 \leq L \leq \frac{1-\|A\|_1}{2a}$,

$$\log \mathbb{E}(e^{Lf(X(k))}) = \log(m(L)) \leq \frac{(a\mathbb{E}(g)+b)L^2}{(1 - \|A\|_1 - 2aL)},$$

so we have

$$\mathbb{E}(e^{\theta f(X)} af(X(k))) \leq a\frac{(a\mathbb{E}(g)+b)L}{(1 - \|A\|_1 - 2aL)}m(\theta) + aL^{-1}\theta m'(\theta).$$

By the convenient choice of $L = -\theta$, we obtain that for $0 \geq \theta \geq -\frac{1-\|A\|_1}{2a}$,

$$\mathbb{E}\left(e^{\theta f(X)}(f(X(k)) + f(X))\right) \leq -a\frac{(a\mathbb{E}(g)+b)\theta}{(1-\|A\|_1+2a\theta)}m(\theta),$$

so for $0 \geq \theta \geq -\frac{1-\|A\|_1}{2a}$,

$$\begin{aligned} m'(\theta) &\geq -\left(e^{-\theta}-1\right)\frac{1}{n}\sum_{k=0}^{\infty}\left(\frac{-a}{2}\frac{(a\mathbb{E}(g)+b)\theta}{(1-\|A\|_1+2a\theta)} + a\mathbb{E}(g)+b\right) \\ &\quad \times m(\theta)\left(1-\frac{1}{n}+\frac{1}{n}\|A\|_1\right)^{k/2} \\ &\geq -\left(e^{-\theta}-1\right)\frac{2}{1-\|A\|_1}\left(\frac{-a}{2}\frac{(a\mathbb{E}(g)+b)\theta}{(1-\|A\|_1+2a\theta)} + a\mathbb{E}(g)+b\right)m(\theta), \end{aligned}$$

which implies (5.3.2). Suppose that $0 \geq \theta \geq -\frac{1-\|A\|_1}{4a}$, then $1-\|A\|_1+2a\theta \geq \frac{1-\|A\|_1}{2}$, so

$$m'(\theta) \geq -\left(e^{-\theta}-1\right)\frac{2}{1-\|A\|_1}\left(\frac{1-\|A\|_1-a\theta}{1-\|A\|_1}(a\mathbb{E}(g)+b)\right)m(\theta), \tag{5.6.13}$$

which implies our mgf bound (5.3.2) in Theorem 5.3.1.

We will split the argument for obtaining tail inequalities in Corollary 5.3.2 into into two parts depending on the size of a.

First, let $K := \frac{1-\|A\|_1}{4a}$, then for $0 \geq \theta \geq -K$, $\left(e^{-\theta}-1\right) \leq \frac{e^K-1}{K}\theta$, and $\frac{1-\|A\|_1-a\theta}{1-\|A\|_1} \leq \frac{5}{4}$, so

$$\begin{aligned} m'(\theta) &\geq -\theta \cdot \frac{e^K-1}{K}\frac{1}{1-\|A\|_1}\frac{5}{2}(a\mathbb{E}(g)+b)m(\theta) \\ \log m(\theta) &\leq \theta^2 \cdot \frac{e^K-1}{K}\frac{1}{1-\|A\|_1}\frac{5}{4}(a\mathbb{E}(g)+b) \leq \frac{2}{1-\|A\|_1}(a\mathbb{E}(g)+b)\theta^2, \end{aligned}$$

whenever

$$\frac{e^K - 1}{K} \leq \frac{8}{5}. \tag{5.6.14}$$

Let us denote the unique positive solution of the equation

$$\frac{e^x - 1}{x} = \frac{8}{5} \tag{5.6.15}$$

by K_c. It is easy to see that $K_c = 1/(4a_c)$. For $K \leq K_c$, (5.6.14) holds, thus for $a \geq \frac{1-\|A\|_1}{4K_c} = (1 - \|A\|_1)a_c$, (5.6.14) holds. Using Markov's inequality, we obtain that for $0 < t < \mathbb{E}(g)$, $0 > \theta > -\frac{1-\|A\|_1}{4a}$,

$$\log \mathbb{P}(f(X) \leq -t) \leq \log(m(\theta)) + t\theta \leq \frac{2}{1 - \|A\|_1}(a\mathbb{E}(g) + b)\theta^2 + \theta t,$$

which takes its minimum at

$$\theta_{\min} = -\frac{(1 - \|A\|_1)t}{4(a\mathbb{E}(g) + b)},$$

which satisfies $0 > \theta_{\min} > -\frac{1-\|A\|_1}{4a}$, and thus

$$\log \mathbb{P}(f(X) \leq -t) \leq -\frac{(1 - \|A\|_1)t^2}{8(a\mathbb{E}(g) + b)}.$$

Finally, we need to verify the case when $a < (1 - \|A\|_1)a_c$. Going back to equation (5.6.13), we can write that for $0 > \theta > -\frac{1-\|A\|_1}{4a}$,

$$m'(\theta) \geq -\left(e^{-\theta} - 1\right) \frac{2}{1 - \|A\|_1} \left(\frac{1 - \|A\|_1 - a\theta}{1 - \|A\|_1}(a\mathbb{E}(g) + b) \right) m(\theta),$$

$$\log(m(\theta))' \geq -\left(e^{-\theta}-1\right)\frac{5}{2}\frac{1}{1-\|A\|_1}(a\mathbb{E}(g)+b),$$
$$\log(m(\theta)) \leq \left(e^{-\theta}+\theta-1\right)\frac{5}{2}\frac{1}{1-\|A\|_1}(a\mathbb{E}(g)+b).$$

Let us write $C := \frac{5}{2}\frac{1}{1-\|A\|_1}(a\mathbb{E}(g)+b)$, then by Markov's inequality, we have that for $0 > \theta > -\frac{1-\|A\|_1}{4a}$, $0 < t < \mathbb{E}(g)$,

$$\log(\mathbb{P}(f(X) \leq -t)) \leq \log(m(\theta)) + \theta t \leq \left(e^{-\theta}+\theta-1\right)C + \theta t$$

The minimum of the right hand side is taken at

$$\theta_{\min} = -\log\left(1+\frac{t}{C}\right) \geq -\log\left(1+\frac{2}{5}\cdot\frac{1-\|A\|_1}{a}\right),$$

which satisfies $0 > \theta_{\min} > -\frac{1-\|A\|_1}{4a}$ whenever $a < a_c(1-\|A\|_1)$. Thus, in this case we have

$$\begin{aligned}\log(\mathbb{P}(f(X) \leq -t)) &\leq \left(\frac{t}{C} - \log\left(1+\frac{t}{C}\right)\right)C - \log\left(1+\frac{t}{C}\right)t \\ &= C\left[\frac{t}{C} - \log\left(1+\frac{t}{C}\right)\left(1+\frac{t}{C}\right)\right]\end{aligned}$$

Now we can verify that the function $x \to x-(1+x)\log(1+x)$ is negative for $x > 0$, and

$$x-(1+x)\log(1+x) \leq -\frac{x^2}{2+(2/3)x},$$

so

$$\log(\mathbb{P}(f(X) \leq -t)) \leq -\frac{t^2}{2C+(2/3)t} = -\frac{t^2}{5(a\mathbb{E}(g)+b)/(1-\|A\|_1)+(2/3)t}. \quad \square$$

5.6.3 The convex distance inequality for dependent random variables

In this section, we prove Theorem 5.3.3. Before turning to the proof, we will state some results. We will use Sion's minimax theorem, which states the following (Sion (1958), and Komiya (1988)).

Theorem 5.6.4. *Let $f(x, y)$ denote a function $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that is convex and lower-semicontinuous with respect to x, concave and upper-semicontinuous with respect to y. If $\mathcal{X}$ is convex and compact, then*

$$\inf_x \sup_y f(x, y) = \sup_y \inf_x f(x, y) = \min_x \sup_y f(x, y).$$

The following lemma is the $*$-self-bounding analogue of Lemma 1 of Boucheron, Lugosi, and Massart (2009).

Lemma 5.6.5. *For any $S \in \mathcal{F}$, $d_T^2(x, S)$ is weakly $(4, 0)$-$*$-self-bounding, and satisfies that $|d_T^2(x, S) - d_T^2(x^*, S)| \leq 1$ for every $x, x^* \in \Lambda$ differing only in one coordinate.*

Proof. The second claim is proven in Lemma 1 of Boucheron, Lugosi, and Massart (2009). The proof of the first claim is similar to the proof of Lemma 1 of Boucheron, Lugosi, and Massart (2009) (see also Proposition 13 of Boucheron, Lugosi, and Massart (2003)). We recall some of their argument here.

Let $\mathcal{M}(S)$ denote the set of probability measures on S. Then, using Sion's minimax theorem, we may rewrite d_T as

$$d_T(x, S) = \inf_{\nu \in \mathcal{M}(S)} \sup_{\|\alpha\|_2 \leq 1} \sum_{j=1}^{n} \alpha_j \mathbb{E}_\nu[\mathbb{1}_{x_j \neq Y_j}] \tag{5.6.16}$$

where $Y = (Y_1, \ldots, Y_n)$ is distributed according to ν.

We may use once again Sion's minimax theorem to write the convex distance as

$$\begin{aligned} d_T(x, S) &= \inf_{\nu \in \mathcal{M}(S)} \sup_{\|\alpha\|_2 \leq 1} \sum_{j=1}^{n} \alpha_j \mathbb{E}_\nu[\mathbb{1}_{x_j \neq Y_j}] \\ &= \sup_{\|\alpha\|_2 \leq 1} \inf_{\nu \in \mathcal{M}(S)} \sum_{j=1}^{n} \alpha_j \mathbb{E}_\nu[\mathbb{1}_{x_j \neq Y_j}]. \end{aligned}$$

Denote the pair (ν, α) at which the saddle point is achieved by $(\hat{\nu}, \hat{\alpha})$.

Note that strictly speaking, the conditions of Sion's minimax theorem ($\mathcal{X}$ should be convex and compact) are not satisfied, however, this problem can be dealt with the same way as in Boucheron, Lugosi, and Massart (2003) (by mapping the large space $\mathcal{M}(S)$ on the convex compact set of the probability measures on $\{0, 1\}^n$).

We can suppose without loss of generality that $d_T^2(y, S) \leq d_T^2(x, S)$, thus

$$\begin{aligned} &d_T^2(x, S) - d_T^2(y, S) = (d_T(x, S) - d_T(y, S))(d_T(x, S) + d_T(y, S)) \\ &\leq (d_T(x, S) - d_T(y, S)) 2 d_T(x, S) \leq \sum_{i: x_i \neq y_i} 2 d_T(x, S) \hat{\alpha}_i, \end{aligned}$$

where $\hat{\alpha}_i$ was defined a few lines above. With

$$\alpha_i(x) := 2 d_T(x, S) \hat{\alpha}_i,$$

we have

$$\sum_{i=1}^{n} \alpha_i(x)^2 \leq 4 d_T^2(x, S),$$

so the claim follows. Similarly, analogously to Proposition 13 of Boucheron, Lugosi, and Massart (2003), one can show that $d_T(x, S)$ is weakly $(1, 0)$-$*$-self-bounding. $\square$

Now we are ready to prove the main result of this section.

Proof of Theorem 5.3.3. By Lemma 5.6.5, we can apply Theorem 5.3.1 to $g(x) := d_T^2(x, S)$ with $a = 4$, $b = 0$. From (5.3.2), we obtain for $0 \geq \theta \geq -\frac{1-\|A\|_1}{8}$,

$$(\log m(\theta))' \geq -\left(e^{-\theta} - 1\right) \frac{2}{1 - \|A\|_1} \left(4\mathbb{E}(g) - \theta \frac{8\mathbb{E}(g)}{(1 - \|A\|_1 + 8\theta)}\right).$$

Here $\left(e^{-\theta} - 1\right) \leq (-\theta)\frac{e^{1/8}-1}{1/8}$. Let us define $\theta^* := \frac{\theta}{1-\|A\|_1}$, then the condition $0 \geq \theta \geq -\frac{1-\|A\|_1}{8}$ above is equivalent to $0 \geq \theta^* \geq -1/8$. Under this assumption, we have

$$(\log m(\theta))' \geq \frac{e^{1/8} - 1}{1/8} \theta^* \left(8\mathbb{E}(g) - \theta^* \frac{16\mathbb{E}(g)}{(1 + 8\theta^*)}\right).$$

By integration we obtain that

$$\log m(\theta) \leq \frac{e^{1/8} - 1}{1/8} \mathbb{E}(g) \left(3(\theta^*)^2 + \frac{1}{4}\theta^* - \frac{1}{32}\log(1 + 8\theta^*)\right)(1 - \|A\|_1).$$

Now by applying Markov's inequality, we obtain

$$\begin{aligned}
\log[\mathbb{P}(X \in S)] = \log[\mathbb{P}(g(X) - \mathbb{E}(g) \leq -\mathbb{E}(g))] &\leq m(\theta) + \theta\mathbb{E}(g) \\
&\leq \frac{e^{1/8} - 1}{1/8} \mathbb{E}(g) \left(3(\theta^*)^2 + \frac{1}{4}\theta^* - \frac{1}{32}\log(1 + 8\theta^*)\right)(1 - \|A\|_1) \\
&\quad + (1 - \|A\|_1)\theta^* \mathbb{E}(g).
\end{aligned}$$

In order to minimize this, we solve

$$\frac{e^{1/8} - 1}{1/8} \theta_m^* \left(8\mathbb{E}(g) - \theta_m^* \frac{16\mathbb{E}(g)}{(1 + 8\theta_m^*)}\right) = -\mathbb{E}(g),$$

which has solution $\theta_m^* \approx -0.0806628 > -1/8$, and thus

$$\mathbb{P}(X \in S) \leq \frac{e^{1/8}-1}{1/8}\mathbb{E}(g)\left(3(\theta_m^*)^2 + \frac{1}{4}\theta_m^* - \frac{1}{32}\log(1+8\theta_m^*)\right)(1-\|A\|_1)$$
$$+ \theta_m^*(1-\|A\|_1)\mathbb{E}(g) \leq -\frac{1}{21.345}(1-\|A\|_1)\mathbb{E}(g).$$

On the other hand, by (5.3.1), we have that for $0 \leq \theta \leq (1-\|A\|_1)/8$,

$$\log \mathbb{E}\left[e^{\theta(g(X)-\mathbb{E}(g))}\right] \leq \frac{4\mathbb{E}(g)\theta^2}{(1-\|A\|_1-8\theta)},$$

thus for $\theta = (1-\|A\|_1)/26.1$,

$$\mathbb{P}(X \in S)\mathbb{E}\left[e^{\theta g(X)}\right]$$
$$\leq \exp\left(\mathbb{E}(g)\left(\theta + \frac{4\mathbb{E}(g)\theta^2}{(1-\|A\|_1-8\theta)} - \frac{1}{21.345}(1-\|A\|_1)\right)\right) \leq 1. \ \square$$

Chapter 6

From Stein-type couplings to concentration

6.1 Introduction

Stein couplings were introduced in Chen and Röllin (2010) as follows.

Definition 6.1.1. Let (W, W', G) be a coupling of square integrable random variables. We call (W, W', G) a Stein coupling if

$$\mathbb{E}\{Gf(W') - Gf(W)\} = \mathbb{E}\{Wf(W)\},$$

for all functions for which the expectation exists.

Remark 6.1.2. In this chapter, in every example we will consider bounded random variables W, W', G, and continuous functions f, thus both of the expectations will exist.

Chen and Röllin (2010) proposes a framework explaining the requirements on the coupling that imply that W is close to normal, and shows many examples of such couplings. Our goal here is to use Stein couplings for proving concentration of W around its mean, that is, bounding the probabilities

$$\mathbb{P}(W - \mathbb{E}(W) \geq t), \quad \text{and} \quad \mathbb{P}(W - \mathbb{E}(W) \leq -t)$$

for $t > 0$. We define the moment generating function of W as $m(\theta) := \mathbb{E}(e^{\theta W})$. It is easy to see that $m'(\theta) = \mathbb{E}(We^{\theta W})$ (if both of the expectations exist). The basic idea of this chapter is that we are going to let $f(x) := e^{\theta x}$, and use the definition of Stein couplings to show that

$$m'(\theta) = \mathbb{E}\{Wf(W)\} = \mathbb{E}\{Gf(W') - Gf(W)\} = \mathbb{E}\left\{G\left(e^{\theta W'} - e^{\theta W}\right)\right\}. \tag{6.1.1}$$

This quantity is then further bounded using information about the typical size of G and $W - W'$. From this bound, we obtain concentration inequalities using a standard argument.

We illustrate our approach with three examples, the number of isolated vertices in an Erdős-Rényi random graph, the number of edges in a geometric random graph, and an example about randomly chose large subgraphs of huge fixed graphs.

All of these examples are based on Stein couplings similar to Construction 2A of Chen and Röllin (2010), which we briefly explain here. Let $X_1, \ldots, X_n$ be dependent centered random variables, and denote $W := X_1 + \ldots + X_n$. Let I be uniformly distributed in $[n] = \{1, 2, \ldots, n\}$, and set $G := -nX_I$. Suppose that we can define $W_1', \ldots, W_n'$ such that for every $1 \leq i \leq n$, $\mathbb{E}(X_i | W_i') = 0$ (this is satisfied in particular

if $\mathbb{E}(X_i) = 0$ and X_i is independent of W_i'). Finally, we set $W' := W_I'$. Then it is easy to check that (W, W', G) is a Stein coupling, satisfying $\mathbb{E}(G|W) = -W$ and $\mathbb{E}(G|W') = 0$.

One basic example where such a coupling is possible is the case of locally dependent random variables (that is, every X_i has a neighbourhood $N_i \subset [n]$ such that X_i is independent of $\{X_j\}_{j \in [N] \setminus N_i}$). In this case, we let

$$W := X_1 + \ldots + X_n, \quad W_i' := \sum\nolimits_{j \in [N] \setminus N_i} X_j, \quad W' := W_I', \text{ and } G := -nX_I.$$

From here, we can obtain concentration inequalities via (6.1.1).

Now we briefly review the related literature. There are several examples in the literature that are using Stein-type couplings to obtain concentration inequalities. The first such approach was proposed in Chatterjee (2005) (see also Chatterjee (2007), Chatterjee and Dey (2010)), where exchangeable pairs are used to obtain concentration inequalities. Note that exchangeable pair couplings are a special case of Stein couplings.

Another approach, which is similar to ours, was proposed in Theorem 3.1 of Chatterjee (2012), where a non-exchangeable coupling structure is used, that is a generalisation of the coupling for locally dependent random variables. As an application, an essentially sharp bound is given to the upper tail of triangle counts in an Erdős-Rényi graph. The main theorem of Chatterjee (2012), however, has been optimised for this particular problem, and makes strong assumptions on the coupling, thus it is not applicable to the examples of this chapter. Our goal here is to state theorems using Stein-type couplings that are useful in a wider variety of problems.

Recently, other Stein-type coupling methods has been proposed for proving concentration inequalities. Ghosh and Goldstein (2011) is based on size-biasing, while Goldstein and Islak (2013) uses zero-biasing.

The chapter is organised as follows. In Section 6.2, we prove a concentration inequality for the number of isolated vertices in Erdős-Rényi graphs. After this, Section 6.3 shows concentration for the number of edges in geometric random graphs. Finally, Section 6.4 proves concentration inequalities for subgraph counts in a random subgraph of a fixed graph whose vertices are sampled without replacement. These results are obtained using abstract lemmas that relate Stein couplings to concentration bounds, which may be of independent interest.

6.2 Number of isolated vertices in Erdős-Rényi graphs

Let $G(n,p)$ be an Erdős-Rényi graph, with edges $X := (X_{i,j})_{1\le i<j\le n}$ being i.i.d. Bernoulli random variables with parameter p. Denote the number of its isolated vertices (that is, the vertices with zero incurring edges) by $\mathcal{I}(X)$. Then the following theorem bounds the lower tail of $\mathcal{I}(X)$. Note that the same bound was shown in Ghosh, Goldstein, and Raič (2011) using size biasing.

Theorem 6.2.1. *For any $t \ge 0$, we have*

$$\mathbb{P}(\mathcal{I}(X) \le \mathbb{E}(\mathcal{I}(X)) - t) \le \exp\left(-\frac{t^2}{4n(1-p)^{n-1}}\right). \tag{6.2.1}$$

To prove this theorem, we will use two lemmas. The first lemma is a well-known result about getting concentration bounds from bounds on the moment generating function. The second shows how to get bounds on the moment generating function

under certain assumptions on the Stein coupling.

Lemma 6.2.2. *Let W be a centered random variable with moment generating function $m(\theta)$. Let $C, D \geq 0$, suppose that $m(\theta)$ is finite, and continuously differentiable in $[0, 1/C)$, and satisfies*

$$m'(\theta) \leq C\theta m'(\theta) + D\theta m(\theta). \tag{6.2.2}$$

Then for $0 \leq \theta < 1/C$,

$$\log(m(\theta)) \leq \frac{D\theta^2}{2(1 - C\theta)}, \tag{6.2.3}$$

and for every $t \geq 0$,

$$\mathbb{P}(W \geq t) \leq \exp\left(-\frac{t^2}{2(D + Ct)}\right). \tag{6.2.4}$$

Remark 6.2.3. For the lower tail, equivalent inequalities hold if we assume that

$$m'(\theta) \geq -C\theta m'(\theta) + D\theta m(\theta) \tag{6.2.5}$$

for $\theta \in (-1/C, 0]$.

Proof. The result follows by a standard Markov inequality argument. □

Lemma 6.2.4. *Let (W, W', G) be a Stein coupling. Suppose that $W \geq W'$ almost surely. Then for any $\theta \geq 0$,*

$$m'(\theta) = \mathbb{E}(-G(e^{\theta W} - e^{\theta W'})) \leq \mathbb{E}(\theta G_-(W - W') \cdot e^{\theta W}). \tag{6.2.6}$$

Similarly, if $W' \geq W$ almost surely, then for any $\theta \leq 0$,

$$m'(\theta) = \mathbb{E}(-G(e^{\theta W} - e^{\theta W'})) \geq \mathbb{E}(\theta G_+(W' - W) \cdot e^{\theta W}). \tag{6.2.7}$$

Here $G_- := -G \cdot \mathbb{1}[G < 0]$ *and* $G_+ := G \cdot \mathbb{1}[G > 0]$ *denotes the negative, and positive parts of* G*.*

Remark 6.2.5. Note that if $\mathbb{E}(G|W') = 0$, then we can shift W' by a constant and ensure that the conditions of this theorem hold.

Proof of Lemma 6.2.4. Since $\theta(W - W') \geq 0$, we have

$$1 - e^{-\theta(W-W')} \leq \theta(W - W'),$$

thus (6.2.6) follows, and the proof of (6.2.7) is similar. □

Proof of Theorem 6.2.1. It is easy to see that $\mathbb{E}(\mathcal{I}(X)) = n(1-p)^{n-1}$, thus we set $W := \mathcal{I}(X) - n(1-p)^{n-1}$. We define X' by picking a vertex I uniformly from $[n]$, and removing all the edges connected to it. Let

$$W' := \mathcal{I}(X') - n(1-p)^{n-1}, \text{ and } G := -n\mathbb{1}[I \text{ is an isolated vertex}] + n(1-p)^{n-1},$$

then (W, W', G) is a Stein coupling, $\mathbb{E}(G|W') = 0$, and $W' \geq W$ almost surely. From Lemma 6.2.4, we obtain that for $\theta < 0$,

$$m'(\theta) \geq \mathbb{E}(G_+\theta(W' - W)e^{\theta W}) \geq n(1-p)^{n-1}\theta\mathbb{E}((W' - W)e^{\theta W}).$$

Now we are left to bound $\mathbb{E}(W' - W|W)$. In the following paragraph, we will show that for any graph X, $\mathbb{E}(W' - W|W) \leq 2$.

Here $W' - W$ expresses the number of new isolated vertices created by erasing all of the edges of a randomly picked vertex from X. This operation can only create new isolated vertices from those that only had one incurring edge. Such vertices are

organised into groups of two (two vertices are connected to each other and isolated from the rest) or groups of $k \geq 3$ ($k-1$ vertices have their only edge connected to the kth vertex, which we call *root vertex*). Let N_k denote the number of groups of size k, for $3 \leq k \leq n$. Since the total number of vertices n, we must have $\sum_{k\geq 2} kN_k \leq n$.

Now if we pick the vertex I from a group of two, that will create two new isolated vertices. If we pick a root vertex from a group of $k \geq 3$, we create k new isolated vertices, while if we pick any other vertex, we create only one new isolated vertex. Therefore, we have

$$\mathbb{E}(W' - W|X) \leq \frac{2N_2}{n} \cdot 2 + \sum_{k=3}^{n} \left(\frac{N_k}{n} k + \frac{(k-1)N_k}{n} \right) \leq \frac{\sum_{k=2}^{n} 2kN_k}{n} \leq 2.$$

This implies that $\mathbb{E}(W' - W|W) \leq 2$, and by substituting this into our bound on the moment generating function, we obtain that for $\theta \leq 0$,

$$m'(\theta) \geq 2n(1-p)^{n-1}\theta m(\theta).$$

From this, we obtain our concentration bound by Lemma 6.2.2. □

6.3 Edge counts in geometric random graphs

Geometric random graphs are a popular model in stochastic geometry (see Penrose (2003), Section 3 for limit theorems for subgraph counts in such graphs). We define a geometric random graph $\mathrm{Geo}(n, c)$ as follows. Let $\Omega = [0, 1]^2$, and $X_1, \ldots, X_n$ be i.i.d. uniform in Ω. Define the distance function $d : \Omega^2 \to \mathbb{R}_+$ as the torus distance between two points (this assumption is made to avoid edge effects). For some $c > 0$,

we put an edge between two points X_i and X_j if their distance is less than c. We call the resulting graph $\mathrm{Geo}(n, c)$.

Theorem 6.3.1. *Denote by $\mathcal{E}$ the number of edges in the geometric random graph* $\mathrm{Geo}(n, c)$. *Let*

$$C_L := \sqrt{6\pi}nc, \quad D_L := 12(\log(1/c) + nc^2\pi)n^2c^2\pi$$
$$C_U := \max(\sqrt{12\pi}nc, 2n), \quad D_U := 24(\log(1/c) + nc^2\pi)n^2c^2\pi.$$

Then for any $t \geq 0$,

$$\mathbb{P}(\mathcal{E} - \mathbb{E}(\mathcal{E}) \geq t) \leq \exp\left(-\frac{t^2}{2(D_U + C_U t)}\right), \text{ and}$$
$$\mathbb{P}(\mathcal{E} - \mathbb{E}(\mathcal{E}) \leq -t) \leq \exp\left(-\frac{t^2}{2(D_L + C_L t)}\right).$$

Remark 6.3.2. Applying McDiarmid's bounded differences inequalities would only give a concentration inequality of order $\exp(-t^2/n^3)$, independent of c. Our result depends on c, thus it is better when c is much smaller than 1.

The proof uses the following two lemmas. The first is a technical result for upper bounding quantities of the form $\mathbb{E}(e^{\theta W}V)$, while the second lemma for obtains moment generating function bounds under certain conditions on the Stein coupling.

Lemma 6.3.3 (Massart (2000))**.** *For real valued random variables V and W, any $L > 0$, for every $\theta \in \mathbb{R}$, we have*

$$\mathbb{E}(e^{\theta W}V) \leq L^{-1}\log\mathbb{E}(e^{LV})m(\theta) + L^{-1}\theta m'(\theta) - L^{-1}m(\theta)\log(m(\theta)),$$

if the expectations on both sides exist.

Proof. Let $U := e^{\theta W}/m(\theta)$. Let $A, B \geq 0$ be two random variables with finite variance and $\mathbb{E}(A) = 1$, then

$$E(A \log(B)) \leq \log(\mathbb{E}(AB))$$

by changing the measure and applying Jensen's inequality. Using this result, we have

$$\begin{aligned}\mathbb{E}(e^{\theta W} V) &= L^{-1} m(\theta) \mathbb{E}\left(U\left(\log \frac{e^{LV}}{U} + \log U\right)\right) \\ &\leq L^{-1} \log \mathbb{E}(e^{LV}) m(\theta) + L^{-1} \mathbb{E}\left(e^{\theta W} \log U\right),\end{aligned}$$

here we have applied our previous inequality with $A = U$ and $B = e^{LV}/U$. Now the result follows using the fact that $\log(U) = \theta W - \log(m(\theta))$. □

Lemma 6.3.4. *Let (W, W', G) be a Stein coupling. Let*

$$G^{(-)} := \text{ess sup}(G) - G, \tag{6.3.1}$$

where ess sup(G) *denotes the supremum of G in the almost sure sense. Suppose that W and W' have the same distribution. Suppose that $W_{\max}$ and $W_{\min}$ are random variables such that $|W - W'| \leq W_{\max} - W_{\min}$, and conditioned on some σ-field $\mathcal{F}$, G is independent of $W_{\max} - W_{\min}$ and W'. Suppose that $W_{\max} - W_{\min} \leq M < \infty$ almost*

surely. Then

$$m'(\theta) \le \mathbb{E}\left(\mathbb{E}\left(G^{(-)}\middle|\mathcal{F}\right)\left(e^{\theta(W_{\max}-W_{\min})}-1\right)e^{\theta W'}\right) \text{ for } \theta>0, \text{ thus} \tag{6.3.2}$$

$$m'(\theta) \le \mathbb{E}\left(2\theta\mathbb{E}\left(G^{(-)}\middle|\mathcal{F}\right)(W_{\max}-W_{\min})e^{\theta W'}\right) \text{ for } 0\le\theta\le 1/M, \text{ and} \tag{6.3.3}$$

$$m'(\theta) \ge \mathbb{E}\left(\theta\mathbb{E}\left(G^{(-)}\middle|\mathcal{F}\right)(W_{\max}-W_{\min})e^{\theta W'}\right) \text{ for } \theta<0. \tag{6.3.4}$$

Proof. For $\theta>0$, using that W and W' have the same distribution, we have

$$\begin{aligned} m'(\theta) &= \mathbb{E}(G(e^{\theta W'}-e^{\theta W})) = \mathbb{E}(G^{(-)}(e^{\theta W}-e^{\theta W'})) \\ &= \mathbb{E}\left(G^{(-)}e^{\theta W'}\left(e^{\theta(W-W')}-1\right)\right) \le \mathbb{E}\left(G^{(-)}\left(e^{\theta(W_{\max}-W_{\min})}-1\right)e^{\theta W'}\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(G^{(-)}\middle|\mathcal{F}\right)\left(e^{\theta(W_{\max}-W_{\min})}-1\right)e^{\theta W'}\right). \end{aligned}$$

The statement for $0<\theta<1/M$ follows from the fact that for $0\le x\le 1$, $e^x-1\le 2x$. For $\theta<0$, using the fact that $1-e^{-x}\le x$ for any $x\in\mathbb{R}$, we have

$$\begin{aligned} m'(\theta) &= \mathbb{E}\left(G^{(-)}e^{\theta W'}\left(e^{\theta(W-W')}-1\right)\right) = -\mathbb{E}\left(G^{(-)}e^{\theta W'}\left(1-e^{\theta(W-W')}\right)\right) \\ &\ge \mathbb{E}(\theta G^{(-)}(W-W')e^{\theta W'}) \ge \mathbb{E}\left(\theta\mathbb{E}\left(G^{(-)}\middle|\mathcal{F}\right)(W_{\max}-W_{\min})e^{\theta W'}\right). \square \end{aligned}$$

Proof of Theorem 6.3.1. Denote by $\mathcal{E}_{i,j}$ the indicator function of the edge between X_i and X_j, then $\mathcal{E}=\sum_{1\le i<j\le n}\mathcal{E}_{i,j}$. We have $\mathbb{E}(\mathcal{E}_{i,j})=c^2\pi$, so $\mathbb{E}(\mathcal{E})=\binom{n}{2}c^2\pi$. Let I and J be random indices such that $I<J$, uniformly chosen among the $\binom{n}{2}$ possibilities. Let $G:=\binom{n}{2}(-\mathcal{E}_{I,J}+c^2\pi)$, then $G^{(-)}=\binom{n}{2}\mathcal{E}_{I,J}$. Define $W=\mathcal{E}-\mathbb{E}(\mathcal{E})$, and W' created by replacing X_I and X_J by an independent copy and evaluating W on the resulting graph. Define $\mathcal{E}_{\max}$ as the maximum number of edges in the geometric random graph that only differs from our graph in X_I and X_J (that is, we move them

to the most dense areas). Similarly, define $\mathcal{E}_{\min}$ as the number of edges of the graph created by removing X_I and X_J. Let $W_{\max} := \mathcal{E}_{\max} - \mathbb{E}(\mathcal{E})$, and $W_{\min} := \mathcal{E}_{\min} - \mathbb{E}(\mathcal{E})$. Then the conditions of Lemma 6.3.4 are satisfied with $\mathcal{F}$ being the σ-field generated by I, J, thus by (6.3.4), for $\theta < 0$,

$$\begin{aligned} m'(\theta) &\geq \mathbb{E}\left(\theta \mathbb{E}\left(G^{(-)} \middle| \mathcal{F}\right)(W_{\max} - W_{\min}) e^{\theta W'}\right) \\ &\geq \theta \binom{n}{2} c^2 \pi \cdot \mathbb{E}\left((W_{\max} - W_{\min}) e^{\theta W'}\right). \end{aligned} \tag{6.3.5}$$

Moreover, we have

$$W_{\max} - W_{\min} \leq 2 \cdot \text{ maximum number of points in a circle of size } c.$$

Now we can cut the square into roughly $1/(4c^2)$ small squares of edge length $2c$, and by putting a circle of radius c into each square, and on the vertices of each square, we cover the original square with roughly $1/(2c^2)$ circles. Since any circle of radius c can cross at most 6 of these circles, we have

$$W_{\max} - W_{\min} \leq 12 \cdot \text{ max. number of points in a circle among the } 1/(2c^2) \text{ circles.}$$

Since the number of points in a circle of radius c is just the sum of n independent Bernoulli random variables with parameter $c^2\pi$, we have that for any $L > 0$,

$$\mathbb{E}\left(e^{L(W_{\max} - W_{\min})}\right) \leq \frac{1}{2c^2}\left(1 - c^2\pi + c^2\pi \cdot e^{12L}\right)^n,$$

and thus

$$\frac{1}{L}\log\mathbb{E}\left(e^{L(W_{\max}-W_{\min})}\right) \leq -\frac{2\log(c)}{L} + \frac{n}{L}\log\left(1-c^2\pi+c^2\pi\cdot e^{12L}\right).$$

From this, by Lemma 6.3.3, and (6.3.5), we have

$$m'(\theta) \geq \\ \theta\binom{n}{2}c^2\pi\cdot\left[\left(-\frac{2\log(c)}{L}+\frac{n}{L}\log\left(1+c^2\pi\cdot\left(e^{12L}-1\right)\right)\right)m(\theta)+L^{-1}\theta m'(\theta)\right].$$

Now with the choice $L = 1/12$, we obtain that for any $\theta < 0$,

$$m'(\theta) \geq \theta\frac{n^2c^2\pi}{2}\cdot\left[24(\log(1/c)+nc^2\pi)m(\theta)+12\theta m'(\theta)\right] = C_1\theta^2m'(\theta)+C_2\theta m(\theta),$$

with $C_1 := 6n^2c^2\pi$ and $C_2 := 12(\log(1/c)+nc^2\pi)n^2c^2\pi$. This bound can be rearranged to obtain that

$$\begin{aligned} m'(\theta)(1-C_1\theta^2) &\geq C_2\theta m(\theta) \\ m'(\theta)(1-\sqrt{C_1}\theta)(1-\sqrt{C_1}\theta) &\geq C_2\theta m(\theta), \\ m'(\theta)(1+\sqrt{C_1}\theta) &\geq \frac{C_2\theta m(\theta)}{1-\sqrt{C_1}\theta} \geq C_2\theta m(\theta) \\ m'(\theta) &\geq -\sqrt{C_1}\theta m'(\theta)+C_2\theta m(\theta). \end{aligned}$$

This means that condition (6.2.5) of Lemma 6.2.2 is satisfied with $C = \sqrt{C_1}$ and $D = C_2$, and the result for the lower tail follows.

For the upper tail, we apply the same argument, but use (6.3.3) of Lemma 6.3.4.

Since we have $W_{\max} - W_{\min} \leq 2n$, thus we obtain that for $0 \leq \theta \leq 1/(2n)$,

$$m'(\theta) \leq 2\theta \binom{n}{2} c^2 \pi \cdot \mathbb{E}\left((W_{\max} - W_{\min}) e^{\theta W'}\right). \tag{6.3.6}$$

Now applying Lemma 6.3.3 with $L = 1/12$ leads to

$$\begin{aligned} m'(\theta) &\leq \theta n^2 c^2 \pi \cdot \left[24(\log(1/c) + nc^2\pi) m(\theta) + 12\theta m'(\theta)\right] \\ &= D_1 \theta^2 m'(\theta) + D_2 \theta m(\theta), \end{aligned}$$

with $D_1 = 12n^2c^2\pi$, and $D_2 = 24(\log(1/c) + nc^2\pi)n^2c^2\pi$. From this, we obtain that for $0 \leq \theta \leq 1/(2n)$,

$$\begin{aligned} m'(\theta)(1 - D_1\theta^2) &\leq D_2 \theta m(\theta) \\ m'(\theta)(1 - \sqrt{D_1}\theta) &\leq D_2 \theta m(\theta) \\ m'(\theta) &\leq D_2 \theta m(\theta) + \sqrt{D_1}\theta m'(\theta) \\ &\leq D_2 \theta m(\theta) + \max(\sqrt{D_1}, 2n)\theta m'(\theta), \end{aligned}$$

thus assumption (6.2.2) is satisfied with $C = \max(\sqrt{D_1}, 2n)$ and $D = D_2$, and the result for the upper tail follows. □

6.4 Large subgraphs of huge graphs

Let us consider a fixed graph with N vertices. Let $[N] := \{1, \ldots, N\}$ denote the vertices of the graph, and and $(E_{i,j})_{1 \leq i < j \leq N}$ denote its edges. We denote the graph by $\mathcal{G} := ([N], (E_{i,j})_{1 \leq i < j \leq N})$.

Now suppose that we choose n vertices out of $[N]$ by sampling without replacement, that is, we let $I(1), \dots, I(n)$ be random variables chosen from $[N]$ such that they are all different, uniformly from the $N \cdot \ldots \cdot (N-n+1)$ possibilities. Let $\mathbb{H} := (\{I(1), \dots, I(n)\}, (E_{I(i),I(j)})_{1 \le i < j \le n})$ be the subgraph of $\mathcal{G}$ with vertices $I(1), \dots, I(n)$.

A natural question is the following. If $\mathcal{F}$ a small fixed subgraph with k vertices, then how many copies of $\mathcal{F}$ are in our subgraph $\mathbb{H}$, and how is this related to the total number of such copies in $\mathcal{G}$? This basically expresses how much can we interfere about the structure of $\mathcal{G}$ from $\mathbb{H}$.

Given a fixed graph $\mathcal{F} := \{[k], (F_{i,j})_{1 \le i < j \le k}\}$, we define the number of induced copies (also called full copies) of $\mathcal{F}$ in $\mathcal{G}$ as

$$N_{\mathcal{F}}(\mathcal{G}) := \sum_{1 \le i(1), \dots, i(k) \le N}{}' \mathbb{1}[E_{i(l),i(m)} = F_{l,m} \text{ for every } 1 \le l < m \le k],$$

where $\sum'$ means that we only add up summands where all the indices are different. Similarly, the number of copies of $\mathcal{F}$ in $\mathcal{G}$ is defined as

$$M_{\mathcal{F}}(\mathcal{G}) := \sum_{1 \le i(1), \dots, i(k) \le N}{}' \mathbb{1}[E_{i(l),i(m)} \ge F_{l,m} \text{ for every } 1 \le l < m \le k].$$

The difference between these two is that the induced copy needs to exactly match $\mathcal{F}$, while a copy only needs to contain all the edges of $\mathcal{F}$ (and can contain more edges). The following theorem expresses that when k is fixed, and both N and n are large, that is, we take large subgraphs of huge graphs, then the number of copies and induced copies of $\mathcal{F}$ in $\mathbb{H}$ is strongly concentrated, and essentially determined by the number of such subgraphs in $\mathcal{G}$.

Theorem 6.4.1. *Let $\mathcal{F} := \{[k], (F_{i,j})_{1 \le i < j \le k}\}$ be a fixed graph with k vertices. Let*

$\mathcal{G} = ([N], (E_{i,j})_{1 \le i < j \le N})$ be a fixed graph with N vertices, and $\mathbb{H}$ be one of its subgraphs with n vertices, chosen uniformly among the $\binom{N}{n}$ possibilities. Denote the number of copies of $\mathcal{F}$ in $\mathcal{G}$ by $N_{\mathcal{F}}(\mathcal{G})$, and the number of copies of $\mathcal{F}$ in $\mathbb{H}$ by $N_{\mathcal{F}}(\mathbb{H})$. Then for any $t \ge 0$, we have

$$\mathbb{P}\left(|N_{\mathcal{F}}(\mathbb{H}) - \mathbb{E}(N_{\mathcal{F}}(\mathbb{H}))| \ge t\right) \le 2\exp\left(-\frac{t^2}{2k^2 n^{k-1} \cdot \mathbb{E}(N_{\mathcal{F}}(\mathbb{H})) + k^2 n^{k-1} t}\right),$$

where $\mathbb{E}(N_{\mathcal{F}}(\mathbb{H})) = N_{\mathcal{F}}(\mathcal{G}) \cdot \frac{n(n-1)\ldots(n-k+1)}{N(N-1)\ldots(N-k+1)}$. The same bounds hold for $M_{\mathcal{F}}(\mathbb{H})$ as well, with $N_{\mathcal{F}}$ replaced by $M_{\mathcal{F}}$ in every formula.

Remark 6.4.2. A weaker bound, of the form

$$\mathbb{P}\left(|N_{\mathcal{F}}(\mathbb{H}) - \mathbb{M}(N_{\mathcal{F}}(\mathbb{H}))| \ge t\right) \le 4\exp\left(-\frac{t^2}{16k^2 n^{2k-1}}\right)$$

can be obtained from equation (6.12) of Theorem 6.5 of Paulin (2012b). Here $\mathbb{M}(N_{\mathcal{F}}(\mathbb{H}))$ denotes the median of $N_{\mathcal{F}}(\mathbb{H})$ (if there are multiple medians, then any of them works).

This theorem can be viewed as a non-asymptotic law of large numbers. When N and n are large, and k is small, and $\mathcal{F}$ is quite frequent in $\mathcal{G}$ in the sense that $N_{\mathcal{F}}(\mathcal{G}) = O(N^k)$, then $\mathbb{E}(N_{\mathcal{F}}(\mathbb{H})) = O(n^k)$, while the typical deviations of $N_{\mathcal{F}}(\mathbb{H})$ is of $O(kn^{k-1/2})$. This implies that $N_{\mathcal{F}}(\mathbb{H})$ is concentrated around its mean, which is determined by $\mathcal{G}$. Thus we can read the structure of $\mathcal{G}$, in the sense of subgraph frequencies, and make small error with high probability, from just one large sample $\mathbb{H}$.

Note that such a similar problem was studied in Tran, Choi, and Zhang (2013), where they count subgraphs in the human genome. However, in contrast with this

chapter, they use sampling with replacement, and only obtain variance bounds, instead of concentration inequalities.

The proof is based on the following lemma.

Lemma 6.4.3. *Let (W, W', G) be a Stein coupling. Suppose that W and W' have the same distribution. Let $G^{(-)}$ be as in* (6.3.1)*. Then*

$$m'(\theta) \leq \mathbb{E}\left\{\theta G^{(-)}|W - W'|\left(\frac{e^{\theta W} + e^{\theta W'}}{2}\right)\right\} \text{ for } \theta > 0, \text{ and}$$
$$m'(\theta) \geq \mathbb{E}\left\{\theta G^{(-)}|W - W'|\left(\frac{e^{\theta W} + e^{\theta W'}}{2}\right)\right\} \text{ for } \theta < 0.$$

Proof. Using the facts that W and W' has the same distribution, and $|e^x - e^y| \leq \frac{e^x+e^y}{2} \cdot |x - y|$ (shown, for example, in Chatterjee (2007)), we have that for $\theta > 0$,

$$\begin{aligned} m'(\theta) &= \mathbb{E}\left\{G\left(e^{\theta W'} - e^{\theta W}\right)\right) = \mathbb{E}\left(G^{(-)}\left(e^{\theta W} - e^{\theta W'}\right)\right\} \\ &\leq \mathbb{E}\left\{G^{(-)}|W - W'|\theta\left(\frac{e^{\theta W} + e^{\theta W'}}{2}\right)\right\}. \end{aligned}$$

The proof for $\theta < 0$ is similar. □

Proof of Theorem 6.4.1. We are going to construct a Stein coupling (W, W', G), and then apply Lemma 6.4.3 to get tail estimates. Note that the construction of this coupling is not in the usual way, since we are going to first define W', then G, and finally W. Although in the statement of the theorem we have already defined $I(1), \ldots, I(n)$ as being sampled without replacement from $[N]$, we will not start the coupling based on this, but later on we will verify that this indeed holds for the construction we make.

Let $p := N_{\mathcal{F}}(\mathcal{G})/[N(N-1)\dots(N-k+1)]$, then

$$\mathbb{E}(N_F(\mathbb{H})) = p \cdot n(n-1)\dots(n-k+1).$$

Let $I'(1),\dots,I'(n)$ be sampled without replacement from $[N]$, and define

$$W' := \sum_{1\le i(1),\dots,i(k)\le n}{}' \left(\mathbb{1}\left[E_{I'(i(l)),I'(i(m))} = F_{l,m} \text{ for every } 1\le l<m\le k\right] - p\right),$$

that is, this is the centered version of the number of copies of $\mathcal{F}$ in the subgraph $\mathbb{H}'$ of $\mathcal{G}$ with vertices $I'(1),\dots,I'(n)$. Let $J(1),J(2),\dots,J(k)$ be sampled without replacement from $[N]$, independently of $I'(1),\dots,I'(n)$, and let

$$G := -n\cdot\ldots\cdot(n-k+1)\cdot\left(\mathbb{1}\left[E_{J(l),J(m)} = F_{l,m} \text{ for every } 1\le l<m\le k\right] - p\right),$$

that is a rescaled, centered version of the indicator function corresponding to whether the subgraph of $\mathcal{G}$ with vertices $J(1),\dots,J(k)$ equals to $\mathcal{F}$.

Now using the independence, we have $\mathbb{E}(G|W') = 0$. We define $I(1),\dots,I(n)$ as follows. First, set $I(1) := I'(1),\dots,I(n) := I'(n)$. Then, whenever an element of the sequence $I(1),\dots,I(n)$ is also a member of the sequence $J(1),\dots,J(k)$, we mark it in both sequences. Suppose that there are r non-marked elements left in the sequence $J(1),\dots,J(k)$. Then we choose r elements at random from the non-marked elements of $I(1),\dots,I(n)$, and replace them with the corresponding non-marked element of $J(1),\dots,J(k)$. This way, we have ensured that the sequence $J(1),\dots,J(k)$

is distributed as if it were sampled without replacement from $I(1), \dots, I(n)$. Let

$$W := \sum_{1 \le i(1), \dots, i(k) \le n}{}' \left(\mathbb{1}\left[E_{I(i(l)), I(i(m))} = F_{l,m} \text{ for every } 1 \le l < m \le k \right] - p \right),$$

then $\mathbb{E}(G|W) = -W$, thus (W, W', G) is a Stein coupling. We can verify that W' and W have the same distribution (actually, they are even exchangeable). Moreover, there are at most k indices i in $[n]$ such that $I(i)$ differs from $I'(i)$, therefore

$$|W - W'| \le n \cdot \ldots \cdot (n-k+1) - (n-k) \cdot \ldots \cdot (n-2k+1) \le k^2 n^{k-1}.$$

Define $G^{(-)} := -G + \mathbb{E}(N_{\mathcal{F}}(\mathbb{H}))$, then $G^{(-)} \ge 0$, and from Lemma 6.4.3, we obtain that for $\theta > 0$,

$$\begin{aligned} m'(\theta) &\le \mathbb{E}\left(\theta G^{(-)} |W - W'| \left(\frac{e^{\theta W} + e^{\theta W'}}{2} \right) \right) \\ &\le \mathbb{E}\left(\theta G^{(-)} \cdot k^2 n^{k-1} \left(\frac{e^{\theta W} + e^{\theta W'}}{2} \right) \right). \end{aligned}$$

Now it is easy to check that $\mathbb{E}(G^{(-)}|W) = W + \mathbb{E}(N_{\mathcal{F}}(\mathbb{H}))$ and $\mathbb{E}(G^{(-)}|W') = \mathbb{E}(N_{\mathcal{F}}(\mathbb{H}))$, thus using the fact that $\mathbb{E}(We^{\theta W}) = m'(\theta)$, we obtain

$$m'(\theta) \le \theta \cdot k^2 n^{k-1} \left(\mathbb{E}(N_{\mathcal{F}}(\mathbb{H})) m(\theta) + m'(\theta)/2 \right). \tag{6.4.1}$$

Now the upper bound follows by applying Lemma 6.2.2 with $D = k^2 n^{k-1} \mathbb{E}(N_{\mathcal{F}}(\mathbb{H}))$ and $C = k^2 n^{k-1}/2$. The lower bound is proven in the same way, except that we use the inequality for $\theta < 0$ in Lemma 6.4.3. Finally, the bounds for $M_{\mathcal{F}}$ can be proven using the same argument. □

Chapter 7

Concentration for local dependence[1]

7.1 Introduction

Local dependence, when the variables only depend on those others which are in their neighborhood, has been one of the first examples of Stein's method, see (Chen and Shao, 2004) and the references therein.

Let $[n] := \{1, \ldots, n\}$. The usual form of local dependence is the following (based on (Chen et al., 2011), Chapter 4.7.).

Definition 7.1.1 ((LD) dependence)**.** A group of random variables $\{X_i\}_{1 \leq i \leq n}$ satisfies (LD) if for each i there exists $A_i \in [n]$, called the *neighbourhood* of X_i, such that X_i and $\{X_j\}_{j \in A_i^c}$ are independent.

Let $\mathcal{G}$ be a graph with vertices $[n]$, and edge between i and j if $i \in A_j$ or $j \in A_i$ (that is, one of them is in the neighborhood of the other). We call $\mathcal{G}$ the *dependency*

[1]This chapter is based on the manuscript Paulin (2012a).

graph.

The *chromatic number* of an undirected graph $\mathcal{G}$, denoted by $\chi(\mathcal{G})$, is the smallest positive integer k such that the vertices of $\mathcal{G}$ can be colored with k colors with no edge between vertices of the same color. An elementary argument shows that $\chi(\mathcal{G})$ is bounded by the maximum degree of the graph $\mathcal{G}$ plus one.

(Janson, 2004) proved concentration of sums under (LD) dependence. In particular, Chernoff-Hoeffding and Bernstein inequalities hold for sums of (LD) dependent variables, with constants less than $\chi(\mathcal{G})$ times weaker than in the independent case. The objective of this chapter is to investigate whether this result holds for more general functions of (LD) dependent variables.

Now we describe the organisation of the chapter. In Section 7.2, via a counterexample, we show that (LD) dependence is a too weak condition for the bounded differences inequality. In Section 7.3, we introduce a stronger condition of local dependence, and show that it implies the bounded differences inequality.

7.2 Counterexample under (LD) dependence

In this section, we show a counterexample illustrating that (LD) dependence is not sufficient for the bounded differences inequality.

Let $n \in \mathbb{N}$ be even. Let $X_1, ..., X_{n/2}$ be i.i.d. Rademacher random variables, with $P(X_i = 1) = P(X_i = -1) = 1/2$. Let Q be an independent Rademacher random variable with $P(Q = 1) = P(Q = -1) = 1/2$. Define $X_{i+n/2} := Q \cdot X_i$ for $1 \leq i \leq n/2$.

Now $\{X_i\}_{1\le i\le n}$ satisfies the (LD) dependence, with

$$A_i := [n] \setminus \{i, n/2+i\} \text{ for } 1 \le i \le n/2, \text{ and}$$
$$A_i := [n] \setminus \{i-n/2, i\} \text{ for } n/2 < i \le n.$$

Now it is easy to see that the dependency graph $\mathcal{G}$ has maximum degree 1, and the chromatic number $\chi(\mathcal{G})$ equals 2. Define the function $g : \{-1,1\}^2 \to \mathbb{R}$ as

$$g(1,1) = g(-1,-1) = 1/2 \text{ and } g(1,-1) = g(-1,1) = -1/2.$$

Let

$$f(x_1, ..., x_n) := \sum_{i=1}^{n/2} g(X_i, X_{i+n/2}) \text{ for } x_1, \ldots, x_n \in \{-1,1\},$$

then f is 1-Hamming Lipschitz in each variable, that is,

$$\sup_{x_1,\ldots,x_n,x_i'} |f(x_1,\ldots,x_i,\ldots,x_n) - f(x_1,\ldots,x_i',\ldots,x_n)| \le 1 \text{ for every } 1 \le i \le n.$$

From the definition of X_i and $X_{i+n/2}$, we can see that $g(X_i, X_{i+n/2}) = Q$ for every $1 \le i \le n/2$, thus

$$f(X_1, ..., X_n) = nQ/2, \tag{7.2.1}$$

taking values $n/2$ and $-n/2$ with probability $1/2$. This behaviour is completely different from the case of independent random variables. If a variant of the bounded differences inequality would hold, then we should have

$$\mathbb{P}(|f(X_1,\ldots,X_n) - \mathbb{E}(f)| \ge t) \le 2\exp(-t^2/(n\cdot c))$$

for some c depending on the chromatic number (or the maximal degree) of the dependency graph $\mathcal{G}$. This is clearly not the case, as (7.2.1) implies that c should be more than $n/8$ in this example, despite the fact that the dependency graph has maximal degree 1.

7.3 Concentration under (HD) dependence

In this section, we define (HD) dependence, a special case of (LD) dependence, and show that it implies the bounded differences inequality.

Definition 7.3.1 ((HD) dependence)**.** We say that random variables $\{X_i\}_{1\le i\le n}$ are *(HD) dependent* if they can be written as functions of independent random variables $\{Y_i\}_{1\le i\le N}$ for some $N \in \mathbb{N}$, that is, there are sets $S_1, \dots, S_n \subset [N]$ and functions $\phi_1, \dots, \phi_n$ such that

$$X_i = \phi_i(Y_{S_i}) \text{ for every } 1 \le i \le n, \text{ where } Y_{S_i} := \{Y_j\}_{j\in S_i}.$$

For each $j \in [N]$, let $R_j := \{i \in [N] \text{ such that } j \in S_i\}$, that is, R_j the set of X_is depending of Y_j, and S_i is the set of Y_js that X_i depends on. We say that $\{X_i\}_{1\le i\le n}$ satisfies *(HD, k, l)* if $\{Y_i\}_{1\le i\le N}$ can be chosen such that

$$\max_{1\le i\le n} |S_i| \le k, \text{ and } \max_{1\le i\le n} |R_i| \le l.$$

The next example illustrates the definition in the case of m-dependence.

Example 7.3.2 (m-dependence)**.** Let $Y_1, \dots, Y_n$ be independent random variables,

and

$$X_1 := f_1(Y_1, \ldots, Y_m), X_2 := f_2(Y_2, \ldots, Y_{m+1}), \ldots, X_n := f_n(Y_n, Y_1, \ldots, Y_{m-1}).$$

A direct application of the definition implies that $X_1, \ldots, X_n$ satisfy (HD, m, m). Moreover, by breaking $(Y_i)_{1 \le i \le n}$ into groups of size m, it follows that $X_1, \ldots, X_n$ also satisfy (HD, 2, $2m-1$).

The next proposition explains the relation between (HD, k, l) and (LD).

Proposition 7.3.3. *(HD,k,l) implies (LD) with a dependency graph $\mathcal{G}$ that has maximum degree bounded by $k(l-1)$.*

Proof. We can choose the neighbourhood A_i of the random variable X_i as the set of the indices of X_js where $S_j \cap S_i$ is non-empty. Since X_i depends on at most k elements of $\{Y_j\}_{1 \le j \le N}$, and each of these influences at most l elements of $\{X_i\}_{1 \le i \le n}$, the size of A_i is bounded by $k(l-1)$. Finally, since the condition that "$S_j \cap S_i$ is non-empty" is symmetric in i and j, it follows that the resulting dependency graph $\mathcal{G}$ has maximum degree at most $k(l-1)$. □

The proposition above implies that the results of (Janson, 2004) also hold for (HD) dependent random variables. Now we show versions of the bounded differences inequality and the method of non-uniformly bounded differences for this dependence structure.

Theorem 7.3.4 (Bounded differences inequality for (HD) dependence)**.** *Suppose that $X = \{X_i\}_{1 \le i \le n}$ satisfies (HD,k,l), $X \in \Lambda$, then for any $f : \Lambda \to \mathbb{R}$ satisfying the*

condition

$$f(x) - f(y) \leq \sum_{1 \leq i \leq n} c_i \mathbb{1}[x_i \neq y_i] \tag{7.3.1}$$

for some $c_1, \ldots, c_n \in \mathbb{R}_+$*, for any* $t \geq 0$*, we have*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2 \exp\left(\frac{-2t^2}{kl \sum_{i=1}^n c_i^2}\right). \tag{7.3.2}$$

Remark 7.3.5. The result is kl times weaker than the bounded differences inequality for independent random variables.

Proof of Theorem 7.3.4. Define $g(Y_1, \ldots, Y_N) := f(\psi_1(Y_{S_1}), \ldots, \psi_n(Y_{S_n})) = f(X)$. For $1 \leq j \leq N$, let $C_j := \sum_{k \in R_j} c_i$, then g satisfies that

$$g(x) - g(y) \leq \sum_{1 \leq j \leq N} c_j \mathbb{1}[x_j \neq y_j] \tag{7.3.3}$$

for any x and y, thus by McDiarmid's bounded differences inequality (see McDiarmid (1989)), we have

$$\mathbb{P}(|g(Y) - \mathbb{E}f(Y)| \geq t) \leq 2 \exp\left(\frac{-2t^2}{\sum_{j=1}^N C_j^2}\right). \tag{7.3.4}$$

Now the result follows by noticing that $\sum_{j=1}^N C_j^2 \leq kl \sum_{i=1}^n c_i^2$. □

Theorem 7.3.6 (Method of non-unif. bounded differences for (HD) dependence)**.** *Suppose that* $X = \{X_i\}_{1 \leq i \leq n}$ *satisfies (HD,k,l),* $X \in \Lambda$*, then for any* $f : \Lambda \to \mathbb{R}$ *satisfying the condition*

$$f(x) - f(y) \leq \sum_{1 \leq i \leq n} c_i(x) \mathbb{1}[x_i \neq y_i] \tag{7.3.5}$$

for some $c_1, \ldots, c_n : \Lambda \to \mathbb{R}_+$ *such that* $\sum_i c_i^2(x) \leq C$ *uniformly, for any* $t \geq 0$*, we have*

$$\mathbb{P}(|f(X) - \mathbb{M}f(X)| \geq t) \leq 4 \exp\left(\frac{-t^2}{4klC}\right), \tag{7.3.6}$$

where $\mathbb{M}f(X)$ *denotes the median of* $f(X)$*.*

Proof. The proof is similar to the proof of the previous theorem. We define $g(Y_1, \ldots, Y_N) = f(X)$ as there, and apply the method of non-uniformly bounded differences (Lemma 6.2.1 on page 122 of Steele (1997)) to $g(Y_1, \ldots, Y_N)$ to conclude. □

Bibliography

R. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:no. 34, 1000–1034, 2008. ISSN 1083-6489.

R. Adamczak and W. Bednorz. Exponential concentration inequalities for additive functionals of markov chains. *arXiv preprint arXiv:1201.3569*, 2012.

R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, 2002a. ISSN 0018-9448.

R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, 2002b. ISSN 0018-9448.

R. Ahlswede and A. Winter. Addendum to: "Strong converse for identification via quantum channels" [IEEE Trans. Inform. Theory **48** (2002), no. 3, 569–579; MR1889969 (2003d:94069)]. *IEEE Trans. Inform. Theory*, 49(1):346, 2003. ISSN 0018-9448.

N. Alon and J. H. Spencer. *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons Inc., Hoboken, NJ,

third edition, 2008. ISBN 978-0-470-17020-5. With an appendix on the life and work of Paul Erdős.

N. Alon, M. Krivelevich, and V. H. Vu. On the concentration of eigenvalues of random symmetric matrices. *Israel J. Math.*, 131:259–267, 2002. ISSN 0021-2172.

C. Ané, S. Blachère, D. Chafaï, P. Fougères, I. Gentil, F. Malrieu, C. Roberto, and G. Scheffer. *Sur les inégalités de Sobolev logarithmiques*, volume 10 of *Panoramas et Synthèses [Panoramas and Syntheses]*. Société Mathématique de France, Paris, 2000. ISBN 2-85629-105-8. With a preface by Dominique Bakry and Michel Ledoux.

D. L. Applegate, R. E. Bixby, V. Chvatal, and W. J. Cook. *The traveling salesman problem: a computational study*. Princeton University Press, 2011.

K. Azuma. Weighted sums of certain dependent random variables. *Tôhoku Math. J. (2)*, 19:357–367, 1967. ISSN 0040-8735.

D. Bakry. Functional inequalities for Markov semigroups. In *Probability measures on groups: recent directions and trends*, pages 91–147. Tata Inst. Fund. Res., Mumbai, 2006.

R. Bardenet and O.-A. Maillard. Concentration inequalities for sampling without replacement. *arXiv preprint arXiv:1309.4029*, 2013.

F. Bauer, J. Jost, and S. Liu. Ollivier-ricci curvature and the spectrum of the normalized graph laplace operator. *arXiv preprint arXiv:1105.3803*, 2011.

F. Bauer, P. Horn, Y. Lin, G. Lippner, D. Mangoubi, and S.-T. Yau. Li-yau inequality on graphs. *arXiv preprint arXiv:1306.2561*, 2013.

S. Bernstein. On a modification of chebyshevs inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

S. Bhamidi, G. Bresler, and A. Sly. Mixing time of exponential random graphs. *Ann. Appl. Probab.*, 21(6):2146–2170, 2011. ISSN 1050-5164.

O. Bormashenko. A coupling argument for the random transposition walk. *arXiv preprint arXiv:1109.3915*, 2011.

S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *Ann. Probab.*, 31(3):1583–1614, 2003. ISSN 0091-1798.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005a. ISSN 1292-8100.

S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560, 2005b. ISSN 0091-1798.

S. Boucheron, G. Lugosi, and P. Massart. On concentration of self-bounding functions. *Electron. J. Probab.*, 14:no. 64, 1884–1899, 2009. ISSN 1083-6489.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013a.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013b.

G. Brightwell and M. Luczak. Extinction times in the subcritical stochastic sis logistic epidemic. *arXiv preprint arXiv:1312.7449*, 2013a.

G. Brightwell and M. Luczak. A fixed-point approximation for a routing model in equilibrium. *arXiv preprint arXiv:1306.5002*, 2013b.

R. Bubley and M. Dyer. Path coupling: A technique for proving rapid mixing in markov chains. In *Foundations of Computer Science, 1997. Proceedings., 38th Annual Symposium on*, pages 223–231. IEEE, 1997.

E. J. Candès and M. A. Davenport. How well can we estimate a sparse vector? *Appl. Comput. Harmon. Anal.*, 34(2):317–323, 2013. ISSN 1063-5203.

S. Chatterjee. *Concentration inequalities with exchangeable pairs*. 2005. ISBN 978-0542-08643-4. Thesis (Ph.D.)–Stanford University, Available at `http://arxiv.org/abs/math.PR/0507526`.

S. Chatterjee. Stein's method for concentration inequalities. *Probab. Theory Related Fields*, 138(1-2):305–321, 2007. ISSN 0178-8051.

S. Chatterjee. The missing log in large deviations for triangle counts. *Random Structures Algorithms*, 40(4):437–451, 2012. ISSN 1042-9832.

S. Chatterjee and P. S. Dey. Applications of Stein's method for concentration inequalities. *Ann. Probab.*, 38(6):2443–2485, 2010. ISSN 0091-1798.

S. Chatterjee and P. Diaconis. Estimating and understanding exponential random graph models. *Ann. Statist.*, 41(5):2428–2461, 2013. ISSN 0090-5364.

S. Chatterjee and Q.-M. Shao. Nonnormal approximation by Stein's method of exchangeable pairs with application to the Curie-Weiss model. *Ann. Appl. Probab.*, 21(2):464–483, 2011. ISSN 1050-5164.

J.-R. Chazottes and F. Redig. Concentration inequalities for Markov processes via coupling. *Electron. J. Probab.*, 14:no. 40, 1162–1180, 2009. ISSN 1083-6489.

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. ISSN 0178-8051.

L. H. Y. Chen and A. Röllin. Stein couplings for normal approximation. *ArXiv e-prints*, Mar. 2010. Available at `http://arxiv.org/pdf/1003.6039`.

L. H. Y. Chen and Q.-M. Shao. Normal approximation under local dependence. *Ann. Probab.*, 32(3A):1985–2028, 2004. ISSN 0091-1798.

L. H. Y. Chen, L. Goldstein, and Q.-M. Shao. *Normal approximation by Stein's method*. Probability and its Applications (New York). Springer, Heidelberg, 2011. ISBN 978-3-642-15006-7.

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics*, 23:493–507, 1952. ISSN 0003-4851.

A. Dembo. Information inequalities and concentration of measure. *Ann. Probab.*, 25 (2):927–939, 1997. ISSN 0091-1798.

L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001. ISBN 0-387-95117-2.

P. Diaconis and L. Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Appl. Probab.*, 6(3):695–750, 1996. ISSN 1050-5164.

P. Diaconis and M. Shahshahani. Generating a random permutation with random transpositions. *Z. Wahrsch. Verw. Gebiete*, 57(2):159–179, 1981. ISSN 0044-3719.

P. Diaconis, E. Mayer-Wolf, O. Zeitouni, and M. P. W. Zerner. The Poisson-Dirichlet law is the unique invariant distribution for uniform split-merge transformations. *Ann. Probab.*, 32(1B):915–938, 2004. ISSN 0091-1798.

P. Diaconis, S. Holmes, and R. Montgomery. Dynamical bias in the coin toss. *SIAM Rev.*, 49(2):211–235, 2007. ISSN 0036-1445.

J. Ding, E. Lubetzky, and Y. Peres. The mixing time evolution of Glauber dynamics for the mean-field Ising model. *Comm. Math. Phys.*, 289(2):725–764, 2009. ISSN 0010-3616.

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. ISSN 0091-1798.

R. L. Dobrushin. Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, 15(3):458–486, 1970.

R. L. Dobrusin. Description of a random field by means of conditional probabilities and conditions for its regularity. *Teor. Verojatnost. i Primenen*, 13:201–229, 1968. ISSN 0040-361x.

W. 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.

R. Douc, E. Moulines, J. Olsson, and R. van Handel. Consistency of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, 39(1):474–513, 2011. ISSN 0090-5364.

D. Dubhashi and D. Ranjan. Balls and bins: a study in negative dependence. *Random Structures Algorithms*, 13(2):99–124, 1998. ISSN 1042-9832.

D. P. Dubhashi and A. Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, Cambridge, 2009. ISBN 978-0-521-88427-3.

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. ISSN 0003-4851.

M. Dyer, L. A. Goldberg, C. Greenhill, M. Jerrum, and M. Mitzenmacher. An extension of path coupling and its application to the Glauber dynamics for graph colourings (extended abstract). In *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 2000)*, pages 616–624, New York, 2000. ACM.

M. Dyer, L. A. Goldberg, and M. Jerrum. Dobrushin conditions and systematic scan. *Combin. Probab. Comput.*, 17(6):761–779, 2008. ISSN 0963-5483.

A. M. Faden. The existence of regular conditional probabilities: necessary and sufficient conditions. *Ann. Probab.*, 13(1):288–298, 1985. ISSN 0091-1798.

D. Fiebig. Mixing properties of a class of Bernoulli-processes. *Trans. Amer. Math. Soc.*, 338(1):479–493, 1993. ISSN 0002-9947.

J. A. 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. ISSN 1050-5164.

D. Gavinsky, S. Lovett, M. Saks, and S. Srinivasan. A tail bound for read-k families of functions. *arXiv preprint arXiv:1205.1478*, 2012.

S. Ghosh and L. Goldstein. Concentration of measures via size-biased couplings. *Probab. Theory Related Fields*, 149(1-2):271–278, 2011. ISSN 0178-8051.

S. Ghosh, L. Goldstein, and M. Raič. Concentration of measure for the number of isolated vertices in the erdős–rényi random graph by size bias couplings. *Statistics & Probability Letters*, 81(11):1565–1570, 2011.

D. Gillman. A Chernoff bound for random walks on expander graphs. *SIAM J. Comput.*, 27(4):1203–1220, 1998. ISSN 0097-5397.

P. W. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic Markov chains. *Statist. Probab. Lett.*, 56(2):143–146, 2002. ISSN 0167-7152.

L. Goldstein and U. Islak. Concentration inequalities via zero bias couplings. *arXiv preprint arXiv:1304.5001*, 2013.

S. Goldstein. Maximal coupling. *Z. Wahrsch. Verw. Gebiete*, 46(2):193–204, 1978/79. ISSN 0044-3719.

N. Gozlan and C. Léonard. Transport inequalities. A survey. *Markov Process. Related Fields*, 16(4):635–736, 2010. ISSN 1024-2953.

G. R. Grimmett and D. R. Stirzaker. *One thousand exercises in probability*. Oxford University Press, 2001. ISBN 9780198572213.

L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975. ISSN 0002-9327.

A. Guillin, C. Léonard, L. Wu, and N. Yao. Transportation-information inequalities for Markov processes. *Probab. Theory Related Fields*, 144(3-4):669–695, 2009. ISSN 0178-8051.

A. Guionnet and B. Zegarlinksi. *Lectures on logarithmic Sobolev inequalities*. Springer, 2003.

B. Gyori and D. Paulin. Non-asymptotic confidence intervals for mcmc in practice. *arXiv preprint*, 2014.

P. Hayden, D. W. Leung, and A. Winter. Aspects of generic entanglement. *Comm. Math. Phys.*, 265(1):95–117, 2006. ISSN 0010-3616.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963. ISSN 0162-1459.

S. Hu. Transportation inequalities for hidden Markov chains and applications. *Sci. China Math.*, 54(5):1027–1042, 2011. ISSN 1674-7283.

F. K. Hwang, D. S. Richards, and P. Winter. *The Steiner tree problem*, volume 53 of *Annals of Discrete Mathematics*. North-Holland Publishing Co., Amsterdam, 1992. ISBN 0-444-89098-X.

S. Janson. Large deviations for sums of partly dependent random variables. *Random Structures Algorithms*, 24(3):234–248, 2004. ISSN 1042-9832.

A. Joulin. A new Poisson-type deviation inequality for Markov jump processes with positive Wasserstein curvature. *Bernoulli*, 15(2):532–549, 2009. ISSN 1350-7265.

A. Joulin and Y. Ollivier. Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.*, 38(6):2418–2442, 2010. ISSN 0091-1798.

N. Kahale. Large deviation bounds for Markov chains. *Combin. Probab. Comput.*, 6 (4):465–474, 1997. ISSN 0963-5483.

S. Karlin and J. McGregor. Many server queueing processes with Poisson input and exponential service times. *Pacific J. Math.*, 8:87–118, 1958. ISSN 0030-8730.

J. H. Kim and V. H. Vu. Concentration of multivariate polynomials and its applications. *Combinatorica*, 20(3):417–434, 2000. ISSN 0209-9683.

H. Komiya. Elementary proof for Sion's minimax theorem. *Kodai Math. J.*, 11(1): 5–7, 1988. ISSN 0386-5991.

A. Kontorovich and R. Weiss. Uniform chernoff and dvoretzky-kiefer-wolfowitz-type inequalities for markov chains and related processes. *arXiv preprint arXiv:1207.4678*, 2012.

L. Kontorovich. Measure concentration of hidden markov processes. *arXiv preprint math/0608064*, 2006.

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`.

L. Kontorovich and K. Ramanan. Concentration inequalities for dependent random variables via the martingale method. *Ann. Probab.*, 36(6):2126–2158, 2008. ISSN 0091-1798.

C. Külske. Concentration inequalities for functions of Gibbs fields with application to diffraction and random Gibbs measures. *Comm. Math. Phys.*, 239(1-2):29–51, 2003. ISSN 0010-3616.

M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Statist.*, 1:63–87 (electronic), 1995/97. ISSN 1292-8100.

M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001. ISBN 0-8218-2864-9.

C. A. León and F. Perron. Optimal Hoeffding bounds for discrete reversible Markov chains. *Ann. Appl. Probab.*, 14(2):958–970, 2004. ISSN 1050-5164.

D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2009. ISBN 978-0-8218-4739-8. With a chapter by James G. Propp and David B. Wilson.

D. A. Levin, M. J. Luczak, and Y. Peres. Glauber dynamics for the mean-field Ising model: cut-off, critical power law, and metastability. *Probab. Theory Related Fields*, 146(1-2):223–265, 2010. ISSN 0178-8051.

P. Lezaud. Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.*, 8(3): 849–867, 1998a. ISSN 1050-5164.

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

P. Lezaud. Chernoff and Berry-Esséen inequalities for Markov processes. *ESAIM Probab. Statist.*, 5:183–201, 2001. ISSN 1292-8100.

T. Lindvall. *Lectures on the coupling method*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1992. ISBN 0-471-54025-0. A Wiley-Interscience Publication.

J. Lott and C. Villani. Ricci curvature for metric-measure spaces via optimal transport. *Ann. of Math. (2)*, 169(3):903–991, 2009. ISSN 0003-486X.

E. Lubetzky and A. Sly. Cutoff for the ising model on the lattice. *Inventiones Mathematicae*, pages 1–37, 2009.

M. Luczak. A quantitative differential equation approximation for a routing model. *arXiv preprint arXiv:1212.3231*, 2012.

M. J. Luczak. Concentration of measure and mixing for Markov chains. In *Fifth Colloquium on Mathematics and Computer Science*, Discrete Math. Theor. Comput. Sci. Proc., AI, pages 95–120. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2008.

M. J. Luczak and C. McDiarmid. Concentration for locally acting permutations. *Discrete Math.*, 265(1-3):159–171, 2003. ISSN 0012-365X.

L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, and J. A. Tropp. Matrix Concentration Inequalities via the Method of Exchangeable Pairs. *ArXiv e-prints*, Jan. 2012.

K. Marton. A measure concentration inequality for contracting Markov chains. *Geom. Funct. Anal.*, 6(3):556–571, 1996a. ISSN 1016-443X.

K. Marton. Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration. *Ann. Probab.*, 24(2):857–866, 1996b. ISSN 0091-1798.

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. ISSN 1016-443X.

K. Marton. Measure concentration for a class of random processes. *Probab. Theory Related Fields*, 110(3):427–439, 1998. ISSN 0178-8051.

K. Marton. Measure concentration and strong mixing. *Studia Sci. Math. Hungar.*, 40(1-2):95–113, 2003. ISSN 0081-6906.

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990. ISSN 0091-1798.

P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab.*, 28(2):863–884, 2000. ISSN 0091-1798.

P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. ISBN 978-3-540-48497-4; 3-540-48497-3. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

B. Maurey. Construction de suites symétriques. *C. R. Acad. Sci. Paris Sér. A-B*, 288 (14):A679–A681, 1979. ISSN 0151-0509.

C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

C. McDiarmid. Concentration for independent permutations. *Combin. Probab. Comput.*, 11(2):163–178, 2002. ISSN 0963-5483.

M. W. Meckes. Concentration of norms and eigenvalues of random matrices. *J. Funct. Anal.*, 211(2):508–524, 2004. ISSN 0022-1236.

E. Milman. Properties of isoperimetric, functional and transport-entropy inequalities via concentration. *Probab. Theory Related Fields*, 152(3-4):475–507, 2012a. ISSN 0178-8051.

E. Milman. Sharp isoperimetric inequalities and model spaces for curvature-dimension-diameter condition. *arXiv preprint arXiv:1108.4609*, 2012b.

M. Mitzenmacher and E. Upfal. *Probability and computing*. Cambridge University Press, Cambridge, 2005. ISBN 0-521-83540-2. Randomized algorithms and probabilistic analysis.

R. Montenegro and P. Tetali. Mathematical aspects of mixing times in Markov chains. *Found. Trends Theor. Comput. Sci.*, 1(3):x+121, 2006. ISSN 1551-305X.

S. V. Nagaev. Some limit theorems for stationary Markov chains. *Teor. Veroyatnost. i Primenen.*, 2:389–416, 1957. ISSN 0040-361x.

Y. Ollivier. Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.*, 256 (3):810–864, 2009. ISSN 0022-1236.

Y. Ollivier. A survey of Ricci curvature for metric spaces and Markov chains. In *Probabilistic approach to geometry*, volume 57 of *Adv. Stud. Pure Math.*, pages 343–381. Math. Soc. Japan, Tokyo, 2010.

Y. Ollivier. A visual introduction to riemannian curvatures and some discrete generalizations. *Analysis and Geometry of Metric Measure Spaces: Lecture Notes of the 50th Seminaire De Mathematiques Superieures (Sms), Montreal, 2011*, 56:197, 2013.

F. Otto and C. Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.*, 173(2):361–400, 2000. ISSN 0022-1236.

D. Paulin. Concentration inequalities in locally dependent spaces. *arXiv preprint*, 2012a.

D. Paulin. Concentration of self-bounding functions in weakly dependent spaces by stein's method. *arXiv preprint*, 2012b.

D. Paulin. Mixing and concentration by Ricci curvature. *arXiv preprint*, 2013.

D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *arXiv preprint*, 2014.

D. Paulin. The convex distance inequality for dependent random variables, with applications to the stochastic travelling salesman and other problems. *ArXiv e-prints*, Jan. 2014.

D. Paulin, L. Mackey, and J. A. Tropp. Deriving matrix concentration inequalities from kernel couplings. *arXiv preprint arXiv:1305.0612*, 2013.

M. Penrose. *Random geometric graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, Oxford, 2003. ISBN 0-19-850626-0.

N. S. Pillai and A. Smith. Finite sample properties of adaptive markov chains via curvature. *arXiv preprint arXiv:1309.6699*, 2013.

E. Rio. Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(10):905–908, 2000. ISSN 0764-4442.

G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probab. Surv.*, 1:20–71, 2004. ISSN 1549-5787.

A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930, 2011. ISSN 0090-5364.

J. S. Rosenthal. Faithful couplings of Markov chains: now equals forever. *Adv. in Appl. Math.*, 18(3):372–381, 1997. ISSN 0196-8858.

L. Saloff-Coste. Lectures on finite Markov chains. In *Lectures on probability theory and statistics (Saint-Flour, 1996)*, volume 1665 of *Lecture Notes in Math.*, pages 301–413. Springer, Berlin, 1997.

P.-M. Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000. ISSN 0091-1798.

M. Sion. On general minimax theorems. *Pacific J. Math.*, 8:171–176, 1958. ISSN 0030-8730.

J. M. Steele. *Probability theory and combinatorial optimization*, volume 69 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997. ISBN 0-89871-380-3.

K.-T. Sturm. On the geometry of metric measure spaces. I. *Acta Math.*, 196(1): 65–131, 2006. ISSN 0001-5962.

M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Inst. Hautes Études Sci. Publ. Math.*, (81):73–205, 1995. ISSN 0073-8301.

M. Talagrand. *Mean field models for spin glasses. Volume I*, volume 54 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer-Verlag, Berlin, 2011. ISBN 978-3-642-15201-6. Basic examples.

N. H. Tran, K. P. Choi, and L. Zhang. Counting motifs in the human interactome. *Nature communications*, 4, 2013.

J. A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Adv. Adapt. Data Anal.*, 3(1-2):115–126, 2011. ISSN 1793-5369.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012. ISSN 1615-3375.

F. Unger. A probabilistic inequality with applications to threshold direct-product theorems. In *Foundations of Computer Science, 2009. FOCS'09. 50th Annual IEEE Symposium on*, pages 221–229. IEEE, 2009.

S. A. van de Geer. On Hoeffding's inequality for dependent random variables. In *Empirical process techniques for dependent data*, pages 161–169. Birkhäuser Boston, Boston, MA, 2002.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

L. Veysseire. A concentration theorem for the equilibrium measure of Markov chains with nonnegative coarse Ricci curvature. *ArXiv e-prints*, Mar. 2012.

L. Veysseire. Coarse ricci curvature for continuous-time markov processes. *arXiv preprint arXiv:1202.0420*, 2012a.

L. Veysseire. *Courbure de Ricci grossière de processus markoviens*. PhD thesis, Ecole normale supérieure de lyon-ENS LYON, 2012b.

C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. Old and new.

V. H. Vu. Concentration of non-Lipschitz functions and applications. *Random Structures Algorithms*, 20(3):262–316, 2002. ISSN 1042-9832. Probabilistic methods in combinatorial optimization.

F. Wang. *Functional Inequalities Markov Semigroups and Spectral Theory*. Elsevier Science, 2006.

N.-Y. Wang. Concentration inequalities for gibbs sampling under the d_{L^2} metric. *Preprint*, 2014.

N.-Y. Wang and L. Wu. Convergence rate and concentration inequalities for gibbs algorithm. *To appear in Bernoulli*, 2014.

L. Wasserman. *All of statistics*. Springer Texts in Statistics. Springer-Verlag, New York, 2004. ISBN 0-387-40272-1. A concise course in statistical inference.

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

L. Wu. Poincaré and transportation inequalities for Gibbs measures under the Dobrushin uniqueness condition. *Ann. Probab.*, 34(5):1960–1989, 2006. ISSN 0091-1798.

K. Yosida. *Functional analysis*, volume 123 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, sixth edition, 1980. ISBN 3-540-10210-8.

Appendices

Appendix A

Concentration for Markov chains

A.1 Counterexample for unbounded sums

In this section, we give a counterexample to a conjecture for concentration of sums of unbounded functions of Markov chains proposed in a previous version of this manuscript.

Lemma 5.5. of Vershynin (2010) shows that three natural definitions of subgaussian random variables (tail bound, moment bound, subexponential moment) are in fact equivalent. Definition 5.7. of Vershynin (2010) defines the ψ_2 norm of a real valued random variable X as

$$||X||_{\psi_2} = \sup_{p \geq 1} p^{-1/2} \left(\mathbb{E}|X|^p\right)^{1/p}. \tag{A.1.1}$$

For bounded variables, we have $||X||_{\psi_2} \leq ||X||_\infty$. Vershynin (2010) states a Chernoff-Hoeffding type inequality for sums of subgaussian random variables.

Proposition A.1.1 (Proposition 5.10 of Vershynin (2010))**.** *Let $X_1, \ldots, X_N$ be independent, centered, subgaussian random variables, and let $K := \max_i ||X_i||_{\psi_2}$. Then for every $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$ and every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} a_i X_i\right| \geq t\right) \leq e \cdot \exp\left(-\frac{ct^2}{K^2 \cdot \sum_{i \leq N} a_i^2}\right), \tag{A.1.2}$$

where $c > 0$ is an absolute constant.

Conjecture (Unbounded random variables)**.** *A version of Proposition 5.10 of Vershynin (2010) holds for Markov chains, with constants t_{mix} times weaker than in the independent case.*

Remark A.1.2. Theorem A.7.1 of Talagrand (2011) an unbounded version of Bernstein's inequality for random variables with exponential tails. See Adamczak (2008) has shown Bernstein-type results for unbounded summands for Markov chains, using regeneration-type assumptions (with additional logarithmic factors).

Here we show this conjecture is false in general. Let $\Omega = \mathbb{R}$, π be the distribution with tails $\pi([x, \infty)) = \pi((-\infty, -x]) = (1/2) \cdot \exp(-x^2)$ for $x \geq 0$, and let $f(x) = x$. Define the operator $\boldsymbol{\pi}$ on $L^2(\pi)$ as $\boldsymbol{\pi}(g)(x) = \mathbb{E}_\pi(g)$, and let $\boldsymbol{P} = \gamma\pi + (1-\gamma)\boldsymbol{I}$ for some $0 < \gamma < 1$. Then this operator $\boldsymbol{P}$ corresponds to a Markov transition kernel P that does the following: in step i (from X_i to X_{i+1}), with probability γ, we set X_{i+1} as an independent variable with distribution π, and with probability $1-\gamma$, $X_{i+1} = X_i$. Then for such a probability transition kernel, it is easy to see that the chain is reversible, with spectral gap γ, and mixing time $t_{\text{mix}} \leq \lceil \log(1/4)/\log(1-\gamma) \rceil) \leq 1 + \log(4)/\gamma$. On the other hand, with probability at least $(1-\gamma)^{n-1}$, $X_1 = X_2 = \ldots = X_n$, so for

every $t \geq 0$,

$$\mathbb{P}_\pi \left(\sum_{i=1}^n f(X_i) \geq t \right) \geq (1-\gamma)^{n-1} \mathbb{P}(f(X_1) \geq t/n) = (1-\gamma)^{n-1} \cdot \frac{1}{2} \exp(-t^2/n^2). \tag{A.1.3}$$

Now with the choice $\gamma = 1/2$, we obtain

$$\mathbb{P}_\pi \left(\sum_{i=1}^n f(X_i) \geq t \right) \geq \exp(-t^2/n^2 - \log(2)n).$$

For large values of t, this is much larger than what we would expect by a Gaussian bound of type (A.1.2). Similarly, for the exponential tail case, we can set $\pi([x, \infty)) = \pi((-\infty, -x]) = (1/2) \cdot \exp(-x)$ for $x \geq 0$. Then for every $t \geq 0$,

$$\mathbb{P}_\pi \left(\sum_{i=1}^n f(X_i) \geq t \right) \geq (1-\gamma)^{n-1} \mathbb{P}(f(X_1) \geq t/n) = (1-\gamma)^{n-1} \cdot \frac{1}{2} \exp(-t/n), \tag{A.1.4}$$

thus for $\gamma = 1/2$, we obtain $\mathbb{P}_\pi \left(\sum_{i=1}^n f(X_i) \geq t \right) \geq \exp(-t/n - \log(2)n)$. For large values of t, this is again much worse than what we would have in the independent case. Thus Conjecture A.1 is false. A possible way to prove inequalities for unbounded summands is truncation (see Propositions 3.3.14 and 3.4.18). This allows us to recover Gaussian/exponential tails for sufficiently small deviation t. Note that for the truncation approach, it is important to know the concentration properties of $f(X_i)$ under the stationary distribution π.

A.2 Coin toss data

Below are results of 10000 coin tosses for Example 3.3.25 (1 corresponds to heads, and 0 to tails).

1001001011110011111100111111111100000110100001111111110101111100000010100000110011000000101111000011000

1000000111111000011111111100111100000001001110000000001110001000011111000101010100001100110110010001111

1100000000111110110000000110100000010101000110111111001000000011010001001111101010111101101111001111100

0110000000000001001100000111110100000011001101111101100000001111111001101011011010000100000101111111111

1101001111111000110000000111110000111101001000101100110101100111001011011100111000001111111001001110011

1011111110011101100000111110010100011001011111110111001001111111111000000000000100000000010110100110000

0101111100000110000011011100101111111001011111100100111111101101110001111111001011100000111111011111000 1

1111000001011001111110001111001111111110000000000111100000011011111111101111000011100111101000111111100

1100011111111000000100101000100100111100000001001101110111000000100011111110000001100011000111100111110

0111011111110010101000011011011001111111100000011111111101000000001110110001110111010111110001001101111

1111111111100011100001111001101010010101011000001101010001000000001100111001110101110101100110001110011

1111100010110011000000000110100001110110011110110101111111111111101111111100000000110000000011111101100

0011010001011001100000000101110001110100011000011100010111100001100000000111000000000110000111111000000

0000010111101111011001100100011111011010111010001111000000000010110010111000001001000011110010001101001

1111111110011001110101011111000000100011111100001110111111111101110110101111000011110000110001000000001

1100000001110000110001011011111011111111010111111111001100111111110011011111000111100100010001111011100

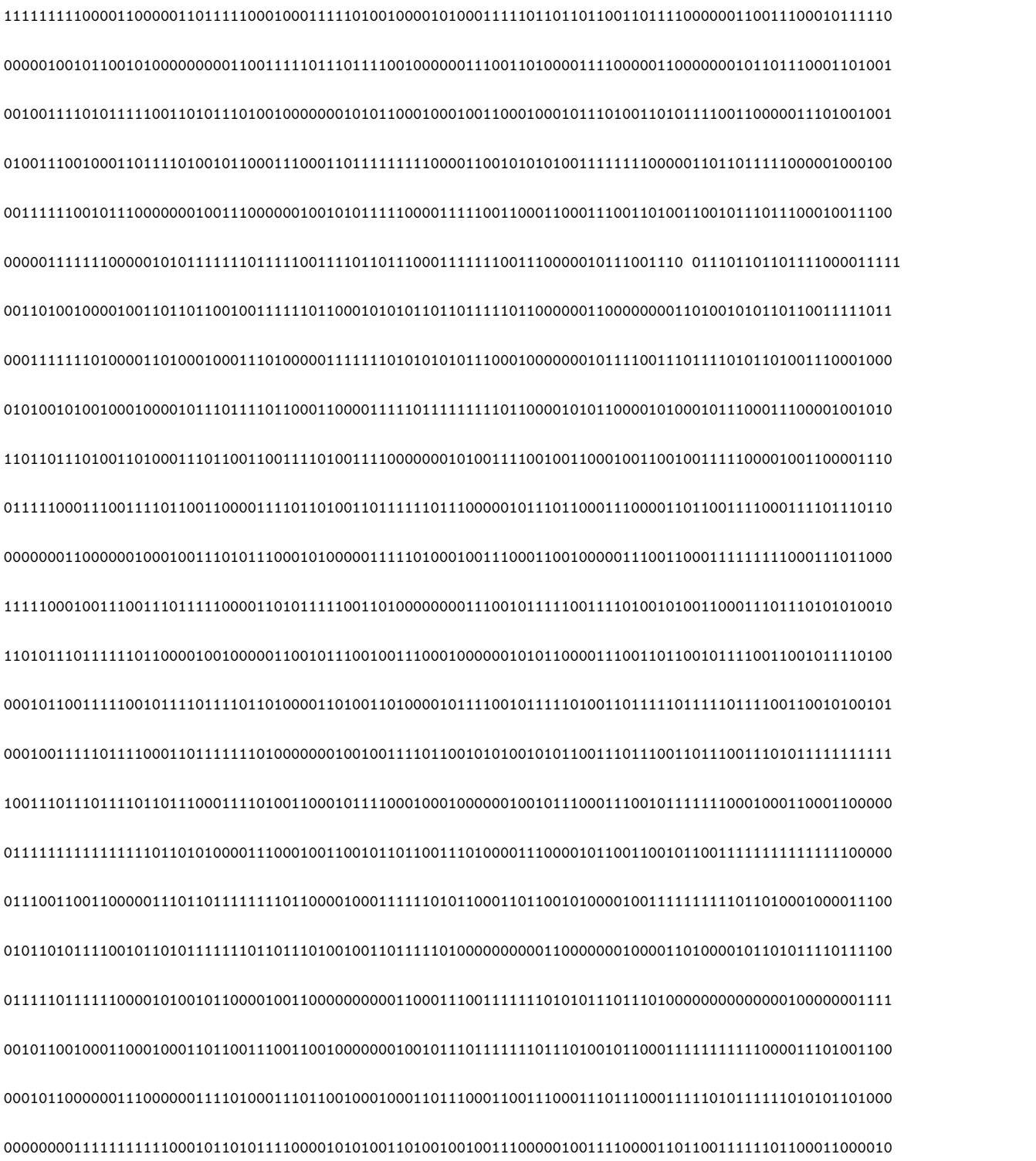
0111101111111110001111000011101001101100000011100000110110001110000000101100001111111100111001111001111

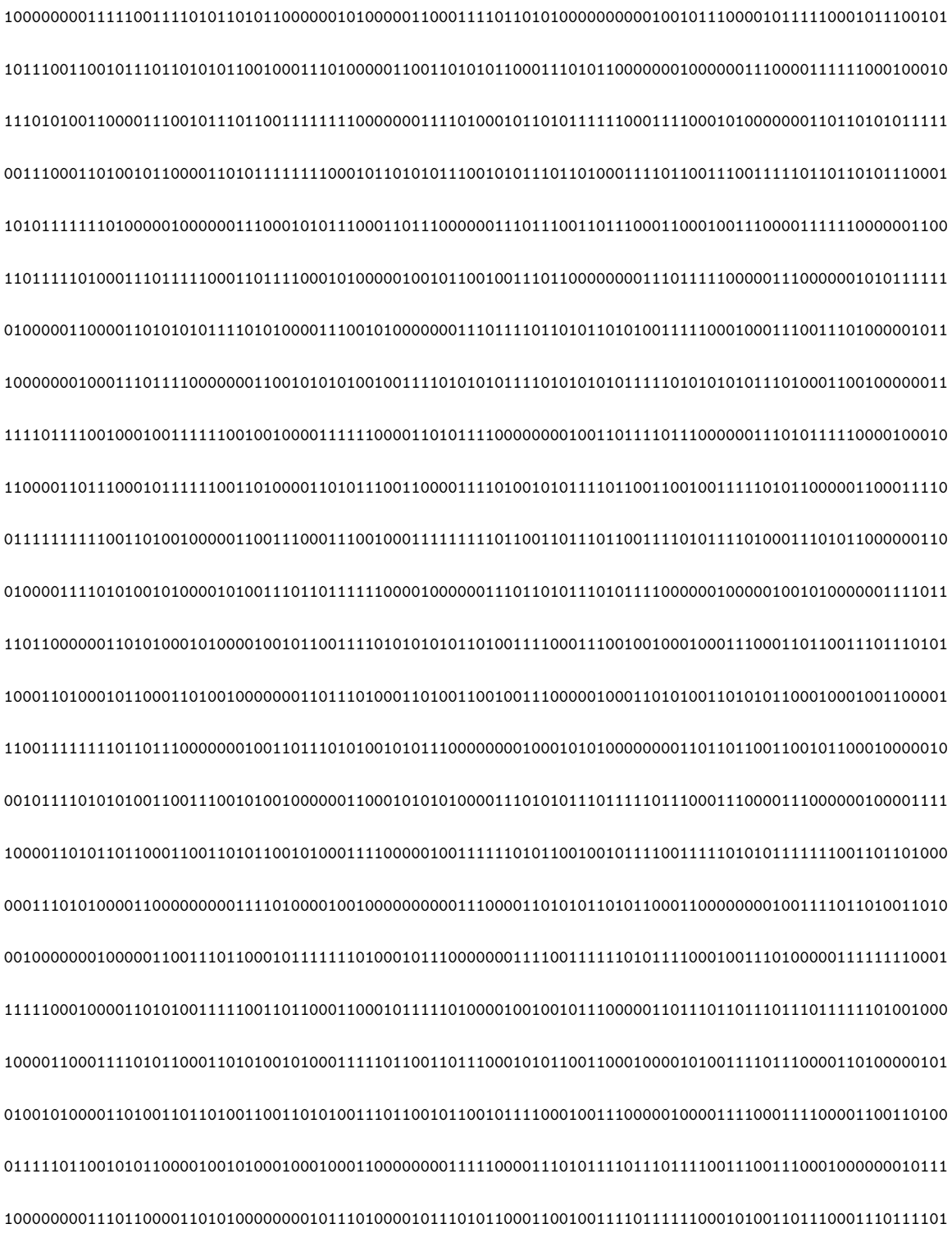
0111011101000110010000110000111100110011100101111111110010101001100000101111100011001011100011101001111

0111100001100011011100100110011111111101001011110000110111111011010000100000010110000000110011100001010111

0000000010111100011011110101001101111000000111011100111001000000000010010100011101111111101110111000 1

0110111001001100011110100000110011111000101000001010110000001110110011111100110100110101000011000110011





1100010011001111111001100001010000101010111000010101001011110000110010101110111111011110000110001001100

1111110011111111101101111111100111111011010000000000110100111001101110000001011010010110110110110110111

1001100000001111001100100100000100101111001011110000111100111110000101001111000000011100111111110001010

1001010001100100000111000010001110110100000011001110101110000010100111111000001011001100000011110010101

0001100111001010100111001011101111000111100111100001111010001111100111100000111101101000000011100011110

1000110011010110010001101100101111101010011110100101100100010011101110010000111100001001111010111111001

0001000010111110010100001000101011011110011111010100100011010111001100101000101010100000111101100110011

0010111111111110101000001101110101011000110110011110000111111100110010010000001110011011011110010000000

0000011001100001011111100000001101111110100100011101000010000100001100000010010001010101001011100011101

1101111101000100010001001011111110101000001101111001100001010111010011011011110111011100010000000000101

1000011010010001101001000110000011100100001010111011010110101100110000000111010111010001110011010011

1011101000000010010101010011111000010001000000000000001111110010011010111001110001001111011000110100110

1000011110101110000000110110010011000011111100101000100001000010101011010010100010100101011011101011 0

1100000001101001010000010101110110000101110110001111000001001011100000011100001110001100000000000100110

1111101110010111001111010001100101000111110001111110001110111010011011001011001100110001101100011111111

0100101100001101111010111110111010011011010010101001101100000110110011011001011010110000100100010010000

1001100001011110100011111101110100011111100000100001001011101100111110111000000011011111101111110000001

0001100000100011111110111011111100001101111111111000111110100011110100000011010111100011111011110011010

0011001111110001111111111111100011001111110110000000100110101011011000100001111000110110001100101011100

1011110001111100110000100101111111001000010001101011010000010110111100111111001110000011010101101110011

1001011101110100011111111001100011011011000001000000011001101011100001010000010010010010011010111000011

0110111111100000000000101110001111110100000100111000011011000011011110011100100000111011111101111100000

0010010010010001001110001001111011100001001000001010001111110000000010110110110 0001111110011010111001111

0111110101110111010001111000011110111000010011110000100001100010011110100000000000000001010100100100111

1001100011101100010111010111011111010010100111101001110100101000110111111101100000011000001110100111010000000110100011010000000111111111000001101001100111100001110001111010011101111010011101001010001101111111011000000110000011101001110100000011010001101000000011111111100000110100110011110000110001111001000000001100111111011111110010000001110111000010110010110100100100100100000111100101000111111111111100001000000111100111100001000101000001101011101111101000011110111101010100000000111110000011000110000111000101110111111010010011111001100111110000101010111001111110000100011111010011101111101010101100001010110110111001010000101110010001110001000011

Appendix B

Convex distance inequality with dependence

B.1 The convex distance inequality for sampling without replacement

In this section, we first state a version of Talagrand's convex distance inequality for sampling without replacement, and then apply it to the stochastic travelling salesmen problem of Section 5.4.1.

Theorem B.1.1. *Let $X = (X_1, \ldots, X_n)$ be a vector of random variables taking values in a set $S = \{A_1, \ldots, A_N\}$. We assume that they are chosen from S without replacement, that is, they are distributed uniformly among the $N \cdot \ldots \cdot (N-n+1)$ possibilities. Let $\Omega := \{x_1, \ldots, x_n \in S, x_i \neq x_j \text{ for } 1 \leq i < j \leq n\}$, then for any $A \subset \Omega$, we have*

$$\mathbb{E}(\exp(d_T^2(X, A)/16)) \leq \frac{1}{\mathbb{P}(A)}, \tag{B.1.1}$$

with d_T defined as in (5.3.4). *Let $g : \Omega \to \mathbb{R}$ be a function satisfying* (5.3.6) *for some functions $c_i : \Omega \to \mathbb{R}_+$, $1 \leq i \leq n$. Suppose that $\sum_{i=1}^n c_i^2(x) \leq C$ for every $x \in \Omega$, then for any $t \geq 0$,*

$$\mathbb{P}(|g(X) - \mathbb{M}(g)| \geq t) \leq 4 \exp\left(\frac{-t^2}{16C}\right), \tag{B.1.2}$$

Remark B.1.2. Note that for sums, Hoeffding and Bernstein-type inequalities for sampling without replacement exist in the literature, see Bardenet and Maillard (2013).

This theorem follows from the following result, due to Talagrand (1995).

Theorem B.1.3. *Denote the symmetric group on $[N]$ by S_N, and let $Y := (Y_1, \ldots, Y_N)$ be distributed uniformly among the $N!$ permutations in S_N. Then for any $B \subset S_N$,*

$$\mathbb{E}(\exp(d_T^2(Y, B)/16)) \leq \frac{1}{\mathbb{P}(B)}.$$

Proof of Theorem B.1.1. Without loss of generality, assume that $S = [N]$. Let us define $B := \{x \in S_N : (x_1, \ldots, x_n) \in A\}$. Then it is easy to check that for this choice, for any $x \in S_N$, $d_T(x, B) = d_T((x_1, \ldots, x_n), A)$. This means that

$$\begin{aligned}\mathbb{E}[\exp(d_T^2((Y_1, \ldots, Y_n), A)/16)] &= \mathbb{E}[\exp(d_T^2(Y, B)/16)] \\ &\leq \frac{1}{\mathbb{P}((X_1, \ldots, X_n) \in B)} = \frac{1}{\mathbb{P}(A)}.\end{aligned}$$

Now (B.1.1) follows from the fact that the vectors $(Y_1, \ldots, Y_n)$ and $(X_1, \ldots, X_n)$ have the same distribution. Finally, we obtain (B.1.2) similarly to the proof of Lemma 6.2.1 on page 122 of Steele (1997). □

As a consequence of these results, we obtain a version of Theorem 5.4.1 for sampling without replacement.

Theorem B.1.4 (Stochastic TSP for sampling without replacement)**.** *Let* $\mathcal{A} = \{a_1, \ldots, a_N\}$ *be a set of points in* $[0,1]^2$*,* $X_1, \ldots, X_n$ *be sampled without replacement from* $\mathcal{A}$*, and* $T(X_1, \ldots, X_n)$ *be the length of the shortest tour according to some cost function* $L(x,y)$ *satisfying* $|x-y| \leq L(x,y) \leq \mathcal{C}|x-y|$ *(as in Section 5.4.1). Then for any* $t \geq 0$*,*

$$\mathbb{P}(|T(X_1, \ldots, X_n) - \mathcal{M}(T)| \geq t) \leq 4 \exp\left(-\frac{t^2}{1024\mathcal{C}^2}\right), \tag{B.1.3}$$

where $\mathcal{M}(T)$ *denotes the median of* T*.*

Proof. This follows from Lemma 5.4.5 and (B.1.2). □