Non-asymptotic confidence intervals for MCMC in practice

Benjamin M. Gyori\(^1\) and Daniel Paulin\(^2\)

\(^{1}\text{NUS Graduate School for Integrative Sciences and Engineering}\)
\(\text{e-mail: bgyori@nus.edu.sg}\)
\(^{2}\text{Department of Mathematics}\)
\(\text{e-mail: paulindani@gmail.com}\)

National University of Singapore
21 Lower Kent Ridge Road, Singapore 119077, Republic of Singapore.

Abstract: Using concentration inequalities, we give non-asymptotic confidence intervals for estimates obtained by Markov chain Monte Carlo (MCMC) simulations, when using the approximation \(\mathbb{E}_\pi f \approx \frac{1}{N} \cdot \sum_{i=1}^{N} f(X_i)\). We state results that are applicable on Markov chains on discrete as well as general state spaces. The confidence intervals are formulated in terms of the spectral properties of the chain, and the properties of \(f\). We also investigate the estimation of \(\mathbb{E}_\pi f\) via subsampling, and by using parallel runs instead of a single run. We propose estimators of the parameters appearing in the bounds, and illustrate our results with simulations on lattice models in statistical physics as well as an example of Bayesian model averaging.

Keywords and phrases: Markov chain Monte Carlo, error bounds, non-asymptotic, confidence interval, concentration inequality, simulation.

AMS 2000 subject classifications: Primary 65C05, 60J10, 62M05; secondary 82B20, 68Q87, 68W20.

1. Introduction

The Monte Carlo method was invented by John von Neumann in the Los Alamos Laboratory, in 1947, for solving the problem of neutron diffusion in fissionable material, and thus helping Edward Teller to build the hydrogen bomb (see Metropolis (1987)).

The Monte Carlo method relies on taking independent samples from a probability distribution to approximate an integral with respect to that distribution. Often, however, it is impossible or impractical to obtain such independent samples. One may still be able to construct a Markov chain whose stationary distribution matches the target distribution. It is then possible to obtain a series of dependent samples by sampling from the Markov chain. This method is called Markov chain Monte Carlo (MCMC).

Let \(X_1, X_2, \ldots\) be a time-homogeneous, ergodic Markov chain, taking values in \(\Omega\), and having stationary distribution \(\pi\). Suppose that we are interested in
computing $E_\pi f$ for some $f : \Omega \to \mathbb{R}$. Then we usually make the approximation

$$E_\pi f \approx \frac{\sum_{i=t_0+1}^N f(X_i)}{N-t_0}, \quad (1.1)$$

for some $t_0 \geq 0$ ("burn-in time"). For $t_0$ fixed, and $N \to \infty$, this average converges to $E_\pi f$ by the ergodic theorem. However, it is not clear how much this convergence is slowed down due to the dependence of the samples. Consequently, an important question in practice is, how large should $N$ be so that this approximation is correct to a certain level of precision? Practitioners often disregard this question, and keep silent about error bounds.

It is well known that the average in (1.1) tends to converge faster for fast mixing chains than for slow mixing ones. But there are few practically applicable results that relate the error in (1.1) to the mixing time and the spectral gap of the chain. In fact, some of the widely used textbooks on Markov chain Monte Carlo (Gilks, Richardson and Spiegelhalter (1996), Rubinstein and Kroese (2008), Robert and Casella (2004)) do not even contain the definition of the mixing time and spectral gap. Researchers in the field generally justify their methods on practical examples, without giving quantitative bounds on the mixing properties of the chain.

Most of the error bounds in the literature are based on asymptotic convergence of the average to the normal distribution. Such asymptotic bounds may underestimate the error for finite sample sizes. In contrast, concentration inequalities can establish non-asymptotic error bounds for MCMC empirical averages of the form (1.1).

Paulin (2014a) has shown the utility of the mixing time and spectral gap for obtaining non-asymptotic error bounds for (1.1). However, when trying to apply these inequalities in practice, we encounter difficulties, since the parameters in the inequalities (such as variance, asymptotic variance, spectral gap, mixing time) are often not known in practice. The main objective of this paper is to show the practical applicability of such concentration bounds, by giving estimators of these parameters based on the data.

We state estimators for the variance and the asymptotic variance ($\text{Var}_\pi(f)$ and $\sigma^2$), and prove concentration inequalities to estimate the precision of these estimators. We state simply computable estimators for the spectral gap and the mixing time, and give a theoretical explanation for their usage.

In addition to the estimators for the parameters, we also show how the concentration bounds of Paulin (2014a) are affected by subsampling (i.e. averaging only every $m$th term in (1.1)), and by running several parallel chains instead of a single chain. We explain when is it worthwhile to use these techniques.

We demonstrate our method on simulation results of various models, including lattice models in statistical physics, where the spectral gap and the mixing time are known, and a case study of Bayesian inference, where the mixing properties of the chain are unknown. In the latter case, we estimate the unknown spectral gap and mixing time from the data.

The structure of the paper is as follows. In Section 1.1, we explain some basics from the theory of general state Markov chains. In the following section
we briefly review the results in the literature about bounding the error of MCMC empirical averages. In Section 3, we state the concentration bounds from Paulin (2014a) in a form that is practically applicable. We also explain how these inequalities are affected by subsampling and parallel runs. In Section 4, we state estimators for the variance, asymptotic variance, spectral gap, and mixing time of the chain. Finally, Section 5 contains our simulation results, and the Appendix contains the proof of the concentration bounds for our estimators on the variance and asymptotic variance.

1.1. Basic definitions for general state Markov chains

In this section, we give some definitions from the theory of general state space Markov chains, based on Roberts and Rosenthal (2004). The total variational distance of two measures $P, Q$ defined on the measurable space $(\Omega, \mathcal{F})$ is defined as

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

which is equivalent to

$$d_{TV}(P, Q) := \inf_{\pi \sim P, Y \sim Q} \mathbb{E}_{\pi}(\mathbb{I}[X \neq Y]),$$

here $\pi \sim P, Y \sim Q$ denotes a distribution on $(\Omega^2, \mathcal{F} \times \mathcal{F})$ with marginals $P$ and $Q$, and the infimum is taken over all such distributions.

We say that a Markov chain is $\phi$-irreducible, if there exists a non-zero $\sigma$-finite measure $\phi$ on $\Omega$ 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 $P^n(x, A) > 0$.

We call a Markov chain with stationary distribution $\pi$ aperiodic if there do not exist $d \geq 2$, and disjoints subsets $\Omega_1, \ldots, \Omega_d \subset \Omega$ with $\pi(\Omega_i) > 0$, $P(x, \Omega_{i+1}) = 1$ for all $x \in \Omega_i$, $1 \leq i \leq d-1$, and $P(x, \Omega_1) = 1$ for all $x \in \Omega_d$.

These properties are sufficient for convergence to a stationary distribution. Theorem 4 of Roberts and Rosenthal (2004) shows that if a Markov chain on a state space with countably generated $\sigma$-algebra is $\phi$-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then for $\pi$-a.e. $x \in \Omega$,

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

Now we define uniform and geometric ergodicity, as in Roberts and Rosenthal (2004). A Markov chain with stationary distribution $\pi$, state space $\Omega$, and transition kernel $P(x, dy)$ is uniformly ergodic if

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

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

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

for some $\rho < 1$, where $M(x) < \infty$ for $\pi$ almost every $x \in \Omega$. 
Remark 1.1. Ergodic Markov chains on finite state spaces are uniformly ergodic. Uniform ergodicity implies \( \phi \)-irreducibility (with \( \phi = \pi \)), and aperiodicity.

The mixing time of a time-homogeneous Markov chain with general state space is defined the following way (see Section 4.5 and 4.6 of Levin, Peres and Wilmer (2009)).

Definition 1.2. Let \( X_1, X_2, X_3, \ldots \) be a time-homogeneous Markov chain with transition kernel \( P(x, dy) \), state space \( \Omega \) (a Polish space), and stationary distribution \( \pi \). The mixing time of the chain, denoted by \( t_{\text{mix}} \), is defined as

\[
d(t) := \sup_{x \in \Omega} d_{\text{TV}} \left( P^t(x, \cdot), \pi \right), \quad \text{and} \quad t_{\text{mix}} := \min \{ t : d(t) \leq 1/4 \}.
\]

The fact that \( t_{\text{mix}} \) is finite is equivalent to the uniform ergodicity of the chain, see Roberts and Rosenthal (2004), Section 3.3.

We call a Markov chain \( X_1, X_2, \ldots \) on state space \( (\Omega, \mathcal{F}) \) with transition kernel \( P(x, dy) \) reversible if there exists a probability measure \( \pi \) on \( (\Omega, \mathcal{F}) \) satisfying the detailed balance conditions,

\[
\pi(dx)P(x, dy) = \pi(dy)P(y, dx) \text{ for every } x, y \in \Omega. \tag{1.4}
\]

Define \( L_2(\pi) \) as the Hilbert space of complex valued measurable functions that are square integrable with respect to \( \pi \), endowed with the inner product \( (f, g) = \int fg^* d\pi \). \( P \) can be then viewed as a linear operator on \( L_2(\pi) \), denoted by \( P \), defined as

\[(Pf)(x) := \mathbb{E}_{P(x, \cdot)}(f),\]

and reversibility is equivalent to the self-adjointness of \( P \). The operator \( P \) acts on measures to the left, i.e. for every measurable subset \( A \) of \( \Omega \),

\[(\mu P)(A) := \int_{x \in \Omega} P(x, A) \mu(dx).\]

For a Markov chain with stationary distribution \( \pi \), we define the spectrum of the chain as

\[S_2 := \{ \lambda \in \mathbb{C} \setminus \{0\} : (\lambda I - P)^{-1} \text{ does not exists as a bounded linear operator on } L_2(\pi) \}.\]

For reversible chains, \( S_2 \) lies on the real line. Now we define the most important spectral properties of the chain, the spectral gap, absolute spectral gap, and the pseudo spectral gap.

Definition 1.3. We define the spectral gap for reversible chains as

\[
\gamma := 1 - \sup \{ \lambda : \lambda \in S_2, \lambda \neq 1 \} \quad \text{if eigenvalue } 1 \text{ has multiplicity } 1,
\]

\[
\gamma := 0 \quad \text{otherwise}.
\]

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

\[
\gamma^* := 1 - \sup \{ |\lambda| : \lambda \in S_2, \lambda \neq 1 \} \quad \text{if eigenvalue } 1 \text{ has multiplicity } 1,
\]

\[
\gamma^* := 0 \quad \text{otherwise}.
\]
For both reversible, and non-reversible chains, we define, as in Paulin (2014a), the pseudo spectral gap, as

$$\gamma_{ps} := \max_{k \geq 1} \left\{ \frac{\gamma((P^*)^k P^k)}{k} \right\},$$

(1.5)

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

**Remark 1.4.** Note that for reversible chains, $\gamma \geq \gamma^*$.

The relation between spectral gap, pseudo spectral gap, and mixing time is given by the following proposition (Propositions 3.3 and 3.4 of Paulin (2014a)).

**Proposition 1.5.** For uniformly ergodic reversible/non-reversible chains, respectively,

$$\gamma^* \geq \frac{1}{1 + \frac{t_{mix}}{\log(2)}}, \quad \gamma_{ps} \geq \frac{1}{2t_{mix}}.$$  

(1.6)

For reversible/non-reversible chains on finite state spaces, respectively,

$$t_{mix} \leq \frac{2\log(2) + \log(1/\pi_{min})}{2\gamma^*}, \quad t_{mix} \leq \frac{1 + 2\log(2) + \log(1/\pi_{min})}{\gamma_{ps}},$$

(1.7)

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

**Remark 1.6.** This proposition means that for Markov chains on finite state spaces, fast mixing and large spectral gap (or pseudo spectral gap) are essentially equivalent.

### 1.2. MCMC sampling schemes

For a random vector with distribution $\pi$, there are many ways to define a Markov chain that has $\pi$ as stationary distribution. Three of the most frequently used are the Metropolis-Hastings chain, the Glauber dynamics with random scan (also called Gibbs sampler and heat-bath dynamics), and the Glauber dynamics with systematic scan. Here we define the most often used variants of these (based on Chapter 3 of Levin, Peres and Wilmer (2009)). Note that some authors mean different things by Glauber dynamics, Gibbs sampler, and heat-bath dynamics.

**Definition** (Metropolis-Hastings chain). Let $\Omega$ be any finite set, and $\Psi$ an irreducible transition matrix. The **Metropolis-Hastings chain** modifies $\Psi$ to obtain a chain with stationary distribution $\pi$. The transition matrix of the Metropolis-Hastings chain for a probability distribution $\pi$ and transition matrix $\Psi$ is defined as

$$P(x, y) = \begin{cases} 
\Psi(x, y) \min \left\{ 1, \frac{\pi(y)\Psi(y, x)}{\pi(x)\Psi(x, y)} \right\} & \text{if } y \neq x, \\
1 - \sum_{z: z \neq x} \Psi(x, z) \min \left\{ 1, \frac{\pi(z)\Psi(z, x)}{\pi(x)\Psi(x, z)} \right\} & \text{if } y = x.
\end{cases}$$

(1.8)
Remark 1.7. An equivalent definition applies in general state spaces with terms \( P(x, dy) \), \( \pi(dx) \) and \( \Psi(x, dy) \) respectively. In most of the practical situations \( \pi(x) = h(x)/Z \), with \( Z \) being a normalization constant that is difficult to determine. A very important feature of the Metropolis-Hastings chain is that the transition probabilities only depend on \( \pi \) through the ratio \( \pi(y)/\pi(x) \), which is independent of \( Z \). The same holds true for the conditional probabilities in the case of the Gibbs sampler.

The Glauber dynamics (with random scan) is a special case of the Metropolis-Hastings chain, when one can directly sample from the conditional distribution of each of the variables given the rest.

Definition (Glauber dynamics). Assume that \( S \) is a Polish space, \( V \) is a finite set of \( S \) valued random variables, \( \Omega = S^{|V|} \), and let \( \pi \) be a distribution on \( \Omega \). Then we define the Glauber dynamics chain as picking one variable in \( V \) uniformly at random, and resampling its value conditionally on the values on the rest of the variables.

Finally, in a frequently used variant of Glauber dynamics, instead of picking the variables at random, they are resampled in sequential order.

Definition (Glauber dynamics with systematic scan). Assume that \( S \) is a Polish space, \( V := (X_1, X_2, \ldots, X_N) \) is a finite set of \( S \) valued random variables, \( \Omega = S^N \), and let \( \pi \) be a distribution on \( \Omega \). Then we define the Glauber dynamics chain with systematic scan as resampling \( X_1, X_2, \ldots, X_N \) in sequential order, each of them conditionally on the values of the rest of the variables.

Remark 1.8. Note that Glauber dynamics with random scan is reversible while Glauber dynamics with systematic scan is non-reversible.

2. Previous results on the error of MCMC averages

In this section we give a brief review of some widely used MCMC convergence diagnostics and error estimation methods. For a more comprehensive overview of available techniques, we refer the reader to Cowles and Carlin (1996), Brooks and Roberts (1998) and Liu (2008).

2.1. Gelman-Rubin diagnostic

One of the most frequently used convergence analysis method is the Gelman-Rubin diagnostic, which was introduced in Gelman and Rubin (1992) (see also Brooks and Gelman (1998), and Gelman et al. (2004)).

The diagnostic uses multiple sequences to assess convergence. First, \( m \) parallel chains are run for \( 2n \) steps. Then the first \( n \) terms of each chain are thrown away. After this, the between, and within sequence variations are computed for each estimated function \( f \) of interest. From these, one computes a quantity called potential scale reduction. Furthermore, one can estimate the effective number of
independent samples, corresponding to the aggregate number of “independent” samples from the $m$ parallel runs.

The chain is assumed to have reached stationary distribution when the potential scale reduction is sufficiently close to 1. The main advantage of this approach is that it is easy to implement, and is available in most statistical packages. However, it does not offer a quantitative bound on the error of the estimate $\frac{1}{n}\sum_{i=1}^{n} f(X_i) - E_\pi f$, and it does not have a rigorous theoretical explanation.

### 2.2. Error estimation by the central limit theorem

The following central limit theorem (CLT) for Markov chains is from Roberts and Rosenthal (2004) (Theorems 23 and 24, and Proposition 29).

**Theorem 2.1.** Let $(X_i)_{i \geq 1}$ be a Markov chain taking values in some general state space $\Omega$, with stationary distribution $\pi$. Let $f : \Omega \to \mathbb{R}$, with $E_\pi (f^2) < \infty$. Define the asymptotic variance of $f$, denoted by $\sigma^2$ as

$$\sigma^2 := \lim_{n \to \infty} \frac{1}{n} \text{Var}_\pi(f(X_1) + \ldots + f(X_n)).$$

(2.1)

Let $Z_n := \frac{1}{n}\sum_{i=1}^{n} f(X_i)$, then if $(X_i)_{i \geq 1}$ is uniformly ergodic, for $\pi$-almost every starting point $x$ (i.e. $X_1 = x$),

$$\sqrt{n}(Z_n - E_\pi f) \Rightarrow N(0, \sigma^2).$$

(2.2)

The same holds if $(X_i)_{i \geq 1}$ is geometrically ergodic, and $E_\pi (f^2 + \delta) < \infty$ for some $\delta > 0$.

**Remark 2.2.** For reversible/non-reversible chains, respectively,

$$\sigma^2 \leq 2\text{Var}_\pi(f)/\gamma, \quad \sigma^2 \leq 4\text{Var}_\pi(f)/\gamma_{\text{ps}},$$

see Paulin (2014a), Theorems 3.5 and 3.7.

To make use of the limiting distribution $N(0, \sigma^2)$, Geyer (1992b) proposes several estimators of $\sigma^2$, including the *initial positive sequence estimator*, and the *initial monotone sequence estimator*. It is proven that under mild conditions on the Markov chain, these estimators asymptotically overestimate $\sigma^2$. However, there is no guarantee that they are close to $\sigma^2$, and we do not know how large sample size is needed.

In Section 4 we will define an estimator of the asymptotic variance, $\hat{\sigma}^2$, and show that it is consistent for bounded functions, and prove non-asymptotic error bounds for it in terms of the mixing time of the chain.

Compared to Bernstein-type concentration inequalities, the CLT approach can be slightly sharper for small deviations from the mean. However, the normal approximation is only true asymptotically, whereas concentration bounds hold for any sample size. Non-asymptotic estimates are available for the quality of the normal approximation (so called Berry-Essèen bounds, see Lezaud (1998b), Lezaud (1998a)). However, the constants in these are too large for practical applicability.
2.3. Concentration inequalities

Here we briefly review the results in the literature about concentration inequalities for empirical averages of Markov chains. Gillman (1998) proved Hoeffding-type concentration inequalities for empirical averages of Markov chains. This was further developed in the seminal work Lezaud (1998b) (see also Lezaud (1998a), Lezaud (2001)), which has shown Bernstein-type inequalities for reversible and non-reversible chains with general state spaces (in the reversible case, they depend on the spectral gap of the chain, while for non-reversible chains, they depend on the spectral gap of the multiplicative reversibilization). A sharp version of the Hoeffding bound for reversible finite state space chains was proven in León and Perron (2004).

In Paulin (2014a), we have shown an improved Bernstein-type inequality for reversible and non-reversible chains (in the reversible case, depending on the asymptotic variance, and the spectral gap of the chain, while in the non-reversible case, on its pseudo spectral gap).

Hoeffding bounds under different, regeneration-type assumptions on the chain were proven in Glynn and Ormoneit (2002), and in Douc et al. (2011). Bernstein inequalities are proven under such assumptions in Adamczak (2008). These are applicable to a larger class of general state space chains than the other results, however, the constants are less explicit, and additional logarithmic terms may appear.

2.4. Error bounds via Ricci curvature

Let \((\Omega, d)\) be a Polish space (metric, complete and separable) and define a Markov chain \(X_1, X_2, \ldots\) on this space with unique stationary distribution \(\pi\). We denote the associated transition kernel \((P_x)_{x \in \Omega}\) so that \(P_x\) is a probability measure on \(\Omega\) and \(P_x(dy)\) is the transition probability from \(x\) to \(y\). The Wasserstein distance of two measures, \(\mu_1\) and \(\mu_2\) on \((\Omega, d)\) is defined as
\[
W_1(\mu_1, \mu_2) := \inf_{\Pi \sim \mu_1 \times \mu_2} \mathbb{E}_\Pi (1_{X \neq Y}),
\]
where the infimum is taken over all couplings of \(\mu_1\) and \(\mu_2\). The coarse Ricci curvature \(\kappa\) is the largest real satisfying that
\[
W_1(P_x, P_y) \leq (1 - \kappa)d(x, y)
\]
for every \(x, y \in \Omega\).

We say that \(f: \Omega \to \mathbb{R}\) is Lipschitz (with respect to \(d\)) if for some \(c \geq 0\), \(|f(x) - f(y)| \leq cd(x, y)\) for every \(x, y \in \Omega\). The smallest \(c \geq 0\) such that this holds is called the Lipschitz coefficient of \(f\), denoted by \(\|f\|_{\text{Lip}}\).

Under the assumption that \(\kappa > 0\), Ollivier (2009) obtains variance and concentration bounds for Lipschitz functions under the stationary distribution \(\pi\). Under the same assumption, Joulin and Ollivier (2010) proves concentration bounds for empirical averages of Lipschitz functions. These bounds depend strongly on the Lipschitz coefficient \(\|f\|_{\text{Lip}}\), the coarse Ricci curvature \(\kappa\), and
on the typical jump size of the Markov chain. Although in some special cases this bound may improve upon our Bernstein-type inequalities, in general, the strong dependence on \( \|f\|_{\text{Lip}} \) ruins the sensitivity of the bound.

### 2.5. Estimation of the mean square error

Various non-asymptotic results also exist for the mean square error of the MCMC estimate. We have included two such bounds in Section 3, Theorem 3.2 and Theorem 3.4. The advantage of these results is that they work even for unbounded functions \( f \) that have a finite variance (with respect to \( \pi \)). Their main disadvantage is that one can only deduce quadratic decay through the Chebyshev inequality instead of exponential or Gaussian decay as with concentration inequalities.

Rudolf (2011) gives a similar bound on the mean square error using the spectral gap \( \gamma \) and \( \|f\|_{p} \) for \( p \geq 2 \). Latuszynski, Miasojedow and Niemiro (2011) gives an asymptotically sharp bound on the mean square error as a function of the asymptotic variance \( \sigma^2 \). The bounds are valid on general state spaces without assuming reversibility, or geometric ergodicity.

### 3. Concentration bounds

In this section, we present concentration inequalities that give non-asymptotic bounds on the approximation (1.1). We state Chebyshev and Bernstein-type inequalities for both reversible and non-reversible chains.

All of the results presented in this section bound the absolute value of the deviation of the estimate from the mean. Due to 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. We make the following assumption in all of the theorems stated in this paper (which we state here to avoid unnecessary repetition).

**Assumption 3.1.** We always assume that the Markov chain \( X_1, \ldots, X_N \) is time homogeneous, \( \phi \)-irreducible, and aperiodic. We also assume that it has a Polish state space \( \Omega \), Markov kernel \( P(x, \text{d}y) \), and denote its unique stationary distribution by \( \pi \).

In the following sections, we will state results about the empirical average defined as

\[
Z := \left( \sum_{i=t_0+1}^{N} f(X_i) \right) / (N - t_0).
\] (3.1)

The quantity \( t_0 \), the so called burn-in time corresponds to the number of samples discarded from the beginning of the chain.

For initial distribution \( q \), and stationary distribution \( \pi \), let

\[
E(t_0) := d_{\text{TV}} \left( qP^{t_0}, \pi \right).
\] (3.2)
This definition will be useful for stating our bounds. For uniformly ergodic Markov chains (including ergodic finite state chains), by Proposition 3.15 of Paulin (2014a), we have

\[ E(t_0) \leq 2^{-\left\lceil n/t_{\text{mix}}\right\rceil}. \] (3.3)

Uniform ergodicity often does not hold for Markov chains on general state spaces. In such cases, we will use a different approach to bound \( E(t_0) \). If a distribution \( q \) on \((\Omega, \mathcal{F})\) is absolutely continuous with respect to \( \pi \), we define the \( \chi^2 \) contrast of \( q \) and \( \pi \), denoted by \( N_q \), as

\[ N_q := \mathbb{E}_\pi \left( \frac{d q}{d \pi} \right)^2 = \int_{x \in \Omega} \frac{d q}{d \pi}(x)q(dx). \] (3.4)

If \( q \) is not absolutely continuous with respect to \( \pi \), then we define \( N_q := \infty \).

When we start from a fixed point \( x \), then \( q(x) = 1 \), and \( N_q = 1/\pi(x) \). Using this quantity, by Proposition 3.15 of Paulin (2014a), for reversible/non-reversible chains, respectively,

\[ E(t_0) \leq \frac{1}{2}(1 - \gamma^*)t_0 \cdot \sqrt{N_q - 1}, \quad E(t_0) \leq \frac{1}{2}(1 - \gamma_{ps})^{(t_0-1)/\gamma_{ps}}/\sqrt{N_q - 1}. \] (3.5)

### 3.1. Reversible chains

First, we state a Chebyshev-type inequality (Theorem 3.5 of Paulin (2014a)).

**Theorem 3.2** (Chebyshev inequality for reversible chains). Let \( X_1, \ldots, X_N \) be a reversible Markov chain with spectral gap \( \gamma \). Let \( f \) be a measurable function from \( \Omega \) to \( \mathbb{R} \), satisfying that \( \mathbb{E}_\pi (f^2) < \infty \). Denote the variance of \( f \) by \( V_f := \text{Var}_\pi(f) \), and let \( \sigma^2 \) be the asymptotic variance of \( f \), defined in (2.1).

If we start from the stationary distribution, then

\[ |\text{Var}_\pi [(f(X_1) + \ldots + f(X_N))/N] - \sigma^2/N| \leq \frac{4V_f}{\gamma^2} \cdot \frac{1}{N^2}. \] (3.6)

Let \( Z \) be as in (3.1). By Chebyshev’s inequality, it follows that for any initial distribution \( q \), for any \( t \geq 0 \), we have

\[ \mathbb{P}_q [|Z - E_q f| \geq t] \leq \frac{\sigma^2 + 4V_f/(N - t_0)\gamma^2}{(N - t_0)t^2} + E(t_0). \] (3.7)

Now we state a Bernstein-type result (Theorem 3.8 of Paulin (2014a)).

**Theorem 3.3** (Bernstein inequality for reversible Markov chains). Let \( X_1, \ldots, X_N \) be a reversible Markov chain with spectral gap \( \gamma > 0 \). Suppose that \( f : \Omega \to \mathbb{R} \) satisfies that \( \sup_{x \in \Omega} |f(x) - \mathbb{E}_\pi f| \leq C \) for some finite \( C \). Let \( V_f \) and \( \sigma^2 \) be defined as in Theorem 3.2, and \( Z \) as in (3.1). Then for any initial distribution \( q \), for any \( t \geq 0 \), we have

\[ \mathbb{P}_q [|Z - E_q f| \geq t] \leq 2 \exp \left[ -\frac{(N - t_0)t^2}{2(\sigma^2 + 0.8V_f + 10tC/\gamma)} \right] + E(t_0). \] (3.8)
3.2. Non-reversible chains

The most popular MCMC methods use reversible chains, in particular, the Metropolis-Hastings algorithm and the Glauber dynamics (with random scan) are reversible. On the other hand, the Glauber dynamics chain with systematic scan is non-reversible. In fact, using non-reversible chains instead of reversible ones can speed up the mixing time in some cases (see Diaconis, Holmes and Neal (2000), and Chen, Lovász and Pak (1999)). Therefore it is of interest to show concentration bounds for non-reversible Markov chains too. The following result is a Chebyshev-type inequality (Theorem 3.7 of Paulin (2014a)).

**Theorem 3.4** (Chebyshev inequality for non-reversible Markov chains). Let $X_1, \ldots, X_N$ be a Markov chain with pseudo spectral gap $\gamma_{\text{ps}}$. Let $f$ be a measurable function from $\Omega$ to $\mathbb{R}$, satisfying that $E_\pi f^2 < \infty$. Let $V_f$ and $\sigma^2$ be defined as in Theorem 3.2. If we start from the stationary distribution, then

$$|\text{Var}_\pi [(f(X_1) + \ldots + f(X_N))/N] - \sigma^2/N| \leq \frac{16V_f}{\gamma_{\text{ps}}^2 \cdot 1/N^2}. \quad (3.9)$$

Let $Z$ be as in (3.1), then by Chebyshev’s inequality, for any initial distribution $q$, for any $t \geq 0$, we have

$$P_q [||Z - E_\pi f|| \geq t] \leq \frac{\sigma^2 + 16V_f/((N - t_0)\gamma_{\text{ps}}^2)}{(N - t_0)t^2} + E(t_0). \quad (3.10)$$

Now we state a Bernstein-type inequality (Theorem 3.10 of Paulin (2014a)).

**Theorem 3.5** (Bernstein inequality for non-reversible Markov chains). Let $X_1, \ldots, X_N$ be a Markov chain with pseudo spectral gap $\gamma_{\text{ps}}$. Suppose that $f : \Omega \to \mathbb{R}$ satisfies that $\sup_{x \in \Omega} |f(x) - E_\pi f| \leq C$ for some finite $C$. Let $V_f$ and $\sigma^2$ be defined as in Theorem 3.2, and $Z$ as in (3.1). Then for any initial distribution $q$, for any $t \geq 0$, we have

$$P_q [||Z - E_\pi f|| \geq t] \leq 2 \exp \left[-\frac{(N - t_0 - 1/\gamma_{\text{ps}}))t^2}{8V_f + 20Ct} \right] + E(t_0). \quad (3.11)$$

**Remark 3.6.** An important assumption of the Bernstein-type inequalities is the boundedness of $f$. Note that via a simple truncation argument, the results can be also extended to unbounded functions, for more details, see Proposition 3.16 of Paulin (2014a).

3.3. Subsampling

$Z := (\sum_{i=t_0+1}^N f(X_i)) / (N - t_0)$ is not the only possible way to approximate $E_\pi f$. We may decide to only average in every $m$th step (typically, we choose $m = 1/\gamma$ for reversible chains, or $m = 1/\gamma_{\text{ps}}$ for non-reversible chains). Assume, without loss of generality, that

$$N = nm \text{ and } t_0 = t_0'm. \quad (3.12)$$
Let $X'_1 := X_m, X'_2 := X_{2m}, \ldots, X'_n := X_{n \cdot m}$, and

$$Z' := \frac{\sum_{i=t'_0+1}^n f(X'_i)}{n-t'_0}. \quad (3.13)$$

Then $X'_1, \ldots, X'_n$ is a Markov chain, which is reversible if the original chain was reversible. In this case, choose $m$ to be odd, then the new transition matrix $P^m$ will have second largest eigenvalue $\lambda^m$ (where $\lambda$ denotes the second largest eigenvalue of $P$), and thus its spectral gap is $\gamma' = 1 - (1 - \gamma)^m$. Let $m_\gamma$ denote the smallest odd number greater or equal to $1/\gamma$, then the choice $m := m_\gamma$ will have pseudo spectral gap $m \cdot \gamma_{ps}$. Therefore, with these choices, almost the same concentration inequalities hold for $Z'$ as for $Z$ (by applying our theorems on $Z'$). The advantage of this approach is that one only needs to compute $f$ in every $1/\gamma$-th (or $m$-th) step, which may result in considerable savings if $f$ is expensive to evaluate.

For further discussion about subsampling, and the optimal choice of the spacing $m$, see Geyer (1992a).

### 3.4. Parallel runs

In this section, first we will show how can we apply the bounds from the previous sections when averaging over several parallel runs instead of a single run.

**Proposition 3.7** (Parallel runs). Suppose that we have $m$ independent parallel chains $X^{(1)}, \ldots, X^{(m)}$ of length $N$, i.e. for $1 \leq i \leq m$, $X^{(i)} = (X_{t'_0}^{(i)}, \ldots, X_N^{(i)})$ are independent time homogenous Markov chains with initial distribution $X_0^{(i)} \sim q$, Polish state space $\Omega$, and stationary distribution $\pi$. We denote by $t_0 \geq 0$ the burn-in time. Let $f : \Omega \to \mathbb{R}$, and define the empirical average

$$Z^{(m)} := \frac{1}{m(N-t_0)} \sum_{i=1}^m \sum_{j=t_0+1}^N f(X_j^{(i)}). \quad (3.14)$$

Then Theorems 3.2, 3.3, 3.4, and 3.5 apply to $Z^{(m)}$ in the place of $Z$ as well, except that we need to replace $N-t_0$ by $m(N-t_0)$, and $E(t_0)$ by $mE(t_0)$ in each case.

**Remark 3.8.** If we choose $t_0$ sufficiently large, then $mE(t_0)$ becomes negligible. In this case, our Proposition says that the bounds for running $m$ chains of length $N$ are equivalent for running a single chain of length $t_0 + m(N-t_0)$. Since typically we should choose $N \gg t_0$, this means that there is almost no difference in the bounds when having a single run or several parallel runs as long as the total number of steps is the same.
Proof. The proof is a simple consequence of the following facts: for independent random variables $Y_1, \ldots, Y_m$, we have $\text{Var}(Y_1 + \ldots + Y_m) = \text{Var}(Y_1) + \ldots + \text{Var}(Y_1)$, and $\mathbb{E} (\exp(\theta(Y_1 + \ldots + Y_m))) = \mathbb{E}(\exp(\theta Y_1)) \cdot \ldots \cdot \mathbb{E}(\exp(\theta Y_m))$. The result follows using these together with the variance and moment generating function estimates (which are given in Paulin (2014a)).

The bounds presented in the previous sections apply to uniformly mixing chains, in particular, to every ergodic Markov chain on finite state spaces. In general state spaces, there are also many situations when our results apply. Namely, in addition to uniformly ergodic chains, they can also be applied to geometrically ergodic reversible chains, as such chains have positive absolute spectral gap $\gamma^*$. In practice, however, there are general state space chains which are not geometrically ergodic, and our bounds are not applicable to these. For example, polynomial ergodicity can occur, see Jarner and Roberts (2002), Jarner and Tweedie (2003). The following proposition gives an alternative estimate to $\mathbb{E}_{\pi} f$ that also works for such situations.

**Proposition 3.9** (Approximately independent empirical averages by parallel runs). Let $M$ be the number of parallel chains. For each $1 \leq I \leq M$, let $X_{0}^{(I)}, X_{1}^{(I)}, \ldots, X_{N}^{(I)}$ be independent Markov chains with Markov kernel $P$, Polish state space $\Omega$, and stationary distribution $\pi$. Assume that $d_{TV}(qP^N, \pi) < \epsilon$ for some $\epsilon \geq 0$. Let $f: \Omega \to \mathbb{R}$ be a measurable function such that $\mathbb{E}_{\pi} f$ exists, then for any $t \geq 0$,

$$
\mathbb{P} \left( \left| \frac{1}{M} \sum_{i=1}^{M} f(X_{N}^{(i)}) - \mathbb{E}_{\pi} f \right| \geq t \right) \leq M\epsilon + \mathbb{P} \left( \left| \frac{1}{M} \sum_{i=1}^{M} f(Y_{N}^{(i)}) - \mathbb{E}_{\pi} f \right| \geq t \right),
$$

where $Y_{1}^{(1)}, \ldots, Y_{N}^{(N)}$ are i.i.d. $\pi$ distributed random variables.

**Remark 3.10.** There are sharp concentration bounds for

$$
\mathbb{P} \left( \left| \frac{1}{N} \sum_{i=1}^{N} f(Y_{i}) - \mathbb{E}_{\pi} f \right| \geq t \right)
$$

in many cases, see the book Boucheron, Lugosi and Massart (2013), and the review Vershynin (2010).

**Remark 3.11.** The moral of this result is that in practice, whenever we estimate $\mathbb{E}_{\pi} f$ for chains that are not geometrically ergodic, short parallel runs can be more reliable than a single long run. Parallel runs can also be computationally advantageous (for example, in the case of parallel computing architectures).

**Proof of Proposition 3.9.** The proof is based on a simple coupling argument. For two probability distributions $\mu$ and $\nu$ on the state space $\Omega$, we can define two random variables $A \sim \mu$ and $B \sim \nu$ on the same probability space such that $\mathbb{P}(A \neq B) = d_{TV}(\mu, \nu)$. This coupling is called the maximal coupling (see Lindvall (1992), Samson (2000) for an explicit construction).
Firstly, given the i.i.d. random variables \( \{X^{(i)}_t\}_{1 \leq i \leq N} \), we define \( Y^{(1)} \sim \pi \) in such a way that \((X^{(1)}, Y^{(1)})\) is maximally coupled, and \(Y^{(1)}\) is independent of all \( \{X^{(i)}_t\}_{2 \leq i \leq N} \). Now suppose that we have already defined \( Y^{(1)}, \ldots, Y^{(i)} \), then we define \( Y^{(i+1)} \sim \pi \) such that it is maximally coupled with \( X^{(i+1)}_t \), and independent of all the other previously defined random variables. Then it is easy to see that for this construction, \( Y^{(1)}, \ldots, Y^{(N)} \) are i.i.d. \( \pi \)-distributed random variables, and \( \mathbb{P}(Y^{(i)} \neq X^{(i)}_t) \leq \epsilon \), thus

\[
\mathbb{P}\left( \left| \frac{1}{N} \sum_{i=1}^{N} f(X^{(i)}_t) - \mathbb{E}_\pi f \right| \geq t \right) \\
\leq \mathbb{P}(Y^{(i)} \neq X^{(i)}_t) \text{ for some } 1 \leq i \leq N + \mathbb{P}\left( \left| \frac{1}{N} \sum_{i=1}^{N} f(Y^{(i)}) - \mathbb{E}_\pi f \right| \geq t \right) \\
\leq N\epsilon + \mathbb{P}\left( \left| \frac{1}{N} \sum_{i=1}^{N} f(Y^{(i)}) - \mathbb{E}_\pi f \right| \geq t \right).
\]

4. Estimation of parameters in practice

The main difficulty we encounter when applying our inequalities is that, in general, we do not know \( V_f = \text{Var}_\pi(f) \) and \( \sigma^2 \) (see (2.1)). In many cases, the spectral gap \( \gamma \), pseudo spectral gap \( \gamma_{ps} \), and mixing time \( t_{\text{mix}} \) are also unknown.

In the next two sections, we are going to give estimates to these quantities based on an initial sample \( f(X_{t_0+1}), \ldots, f(X_{N}) \).

4.1. Estimation of the variance and the asymptotic variance

From the definitions, it is easy to see that we can estimate \( V_f \) as

\[
\hat{V}_f := \frac{1}{N - t_0} \left( \sum_{i=t_0+1}^{N} f^2(X_i) \right) - \left( \frac{1}{N - t_0} \sum_{i=t_0+1}^{N} f(X_i) \right)^2.
\] (4.1)

The next proposition gives a bound on the upper tails of \( V_f - \hat{V}_f \) (the proof can be found in the Appendix).

**Proposition 4.1.** Suppose that \( X_1, \ldots, X_N \) is an uniformly ergodic Markov chain, with stationary distribution \( \pi \), and initial distribution \( q \). For any \( T \geq 0 \),

\[
\mathbb{P}_{q}\left( V_f - \hat{V}_f \geq \frac{8t_{\text{mix}}}{N - t_0} + T \right) \leq \exp\left( \frac{-(N - \hat{t}_0)T^2}{200C^4t_{\text{mix}}} \right) + E(\hat{t}_0).
\] (4.2)
Now we propose an estimator to the asymptotic variance $\sigma^2$. For some integer $k \in [1, N - t_0 - 1]$, let

$$
\hat{\sigma}^2(k) := \left( \gamma_0 + 2 \sum_{i=1}^{k} \hat{\gamma}_i \right) \cdot \frac{N - t_0 + k + 1}{N - t_0 - k},
$$

with

$$
\hat{\gamma}_i := \frac{\sum_{j=i_0 + 1}^{N - k} f(X_j) f(X_{j+i})}{N - t_0 - k} \left( \frac{\sum_{j=i_0 + 1}^{N - k} f(X_j)}{N - t_0 - k} \right)^2 - \frac{1}{2} \left( \frac{\sum_{j=i_0 + 1}^{N - k} f(X_j)}{N - t_0 - k} \right)^2.
$$

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

**Proposition 4.2.** For stationary, reversible chains, when $k$ is even, the expected value of $\hat{\sigma}^2(k)$ satisfies the following inequality:

$$
-L_k \leq \sigma^2 - E_\pi(\hat{\sigma}^2(k)) \leq U_k,
$$

with

$$
L_k := \left( \min \left( V_f, \frac{2 V_f}{\gamma} (1 - \gamma^*)^{k+1} \right) + \frac{4 V_f}{\gamma^2} \frac{2 k + 1}{(N - t_0 - k)^2} \right) \cdot \frac{N - t_0 - k}{N - t_0 - 3k - 1},
$$

and

$$
U_k := \left( \frac{2 V_f}{\gamma} (1 - \min(\gamma, 1))^{k+1} + \frac{4 V_f}{\gamma^2} \frac{2 k + 1}{(N - t_0 - k)^2} \right) \cdot \frac{N - t_0 - k}{N - t_0 - 3k - 1}.
$$

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

$$
|E_\pi(\hat{\sigma}^2(k)) - \sigma^2| \leq W_k,
$$

with

$$
W_k := \frac{4 V_f}{\gamma ps} (1 - \gamma ps)^{(k+1 - 1/\gamma ps)/2} + \frac{16 V_f}{\gamma ps^2} \frac{2 k + 1}{(N - t_0 - k)^2}.
$$

**Proposition 4.3.** Suppose that $f : \Omega \rightarrow \mathbb{R}$ satisfies that $\sup_{x \in \Omega} |f(x) - E_\pi f| \leq C$ for some finite $C$. In the case of stationary, uniformly ergodic chains, we have for any $t \geq 0$,

$$
P_\pi( |\hat{\sigma}^2(k) - E_\pi(\hat{\sigma}^2(k))| \geq t ) \leq 2 \exp \left( -\frac{t^2 (N - t_0 - 3k - 1)}{512 (2k + 1)^2 C_{mix}^4} \right),
$$

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

$$
P_q( \sigma^2 - \hat{\sigma}^2(k) \geq U_k + t ) \leq \exp \left( -\frac{t^2 (N - t_0 - 3k - 1)}{512 (2k + 1)^2 C_{mix}^4} \right) + E(t_0),
$$
and for uniformly ergodic non-reversible chains, for any \( k \geq 1, t \geq 0, \)

\[
P_q (\sigma^2 - \hat{\sigma}^2(k) \geq W_k + t) \leq \exp \left( -\frac{t^2 (\hat{N} - \hat{t}_0 - 3k - 1)}{512(2k+1)^2 C^4 t_{\text{mix}}^2} \right) + E(\hat{t}_0). \tag{4.9}
\]

**Remark 4.4.** 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

\[
\hat{t}_0 := \lfloor 0.1\hat{N} \rfloor, \quad k := 10 \cdot \lfloor \hat{N}^{1/3} \rfloor, \quad \hat{\sigma}^2 := \hat{\sigma}^2(k), \tag{4.10}
\]

our bounds imply that for bounded functions, \( \hat{\sigma}^2 \) will be a consistent estimate of \( \sigma^2 \) as \( \hat{N} \to \infty \), for any uniformly ergodic Markov chain, irrespectively of the value of the mixing time. It is also clear that based on this bound, \( \hat{N} \gg t_{\text{mix}}^3 \) samples are necessary to obtain a good estimate of \( \sigma^2 \). In practice, we suggest choosing \( \hat{N} \) to be at least \( 10^6 \), or higher.

**Remark 4.5.** Note that via a simple truncation argument, the error bound of Proposition 4.3 can be also extended to unbounded functions (similarly to Proposition 3.16 of Paulin (2014a)).

### 4.2. Estimation of the spectral gap and the mixing time

Precise estimation of the spectral gap and the mixing time from the realizations \( f(X_1), \ldots, f(X_N) \) is not possible, since it is a property of the Markov chain \( X_1, \ldots, X_N \) itself, and by applying the function \( f \), we lose information. Nevertheless, in practice, we have found that the simple estimate in (4.12) often works well. We now give a brief justification of this approach.

For reversible chains with state space \( \Omega \), transition matrix \( P \), and stationary distribution \( \pi \), define \( l^2(\pi) \) as the Hilbert space of real valued functions on \( \Omega \), with scalar product

\[
\langle f, g \rangle := \sum_{x \in \Omega} f(x)g(x)\pi(x).
\]

Let \( \{\varphi_i\}_{i \geq 1} \) be an orthonormal basis made of the eigenvectors of \( P \), corresponding to eigenvalues \( (\lambda_i)_{i \geq 1} \). Then the largest eigenvalue \( \lambda_1 = 1 \), and \( \varphi_1 = 1 \), and obviously \( \lambda_2 = 1 - \gamma \). By Proposition 1.5 on page 48 of Lezaud (1998b), we have

\[
\sigma^2 = \sum_{i \geq 2} \langle f, \varphi_i \rangle^2 \frac{1 + \lambda_i}{1 - \lambda_i} \leq \frac{2V_f}{\gamma}, \tag{4.11}
\]

thus \( \gamma \leq 2V_f/\sigma^2 \) for any function \( f \). With the choice \( f = \varphi_1 \), we have \( \sigma^2 = (2 - \gamma)/\gamma \), and \( V_f = 1 \), thus \( 2V_f/\sigma^2 = \gamma \cdot 2/(2 - \gamma) \), which is indeed very close to \( \gamma \) (in practice, usually \( \gamma \ll 1 \)). For this reason, if we are only equipped with the values of a single function \( f, f(X_1), \ldots, f(X_\hat{N}) \), then we propose the estimate

\[
\hat{\gamma} := \frac{2V_f}{\hat{\sigma}^2}. \tag{4.12}
\]
with \( \hat{V}_f \) defined as in (4.1), with \( \hat{t}_0 = [0.1\hat{N}] \), and \( \hat{\sigma}^2 \) defined as in (4.10).

In case we have the values \( f(X_1), \ldots, f(X_N) \) for several functions \( f_1, f_2, \ldots, f_k \), then denote the corresponding estimates of \( \hat{\sigma}^2 \) and \( \hat{V}_f \) by \( \hat{\sigma}^2_1, \ldots, \hat{\sigma}^2_k \) and \( \hat{V}_{f_1}, \ldots, \hat{V}_{f_k} \), and estimate \( \gamma \) as

\[
\hat{\gamma} := \min_{1 \leq i \leq k} \frac{2\hat{V}_{f_i}}{\hat{\sigma}^2_i}.
\]  
(4.13)

Similarly, for non-reversible chains, from Theorem 3.4, and the definition of \( \sigma^2 \), it follows that \( \sigma^2 \leq \frac{4\hat{V}_f}{\gamma_{ps}} \). If the values of a single function \( f \) are available, we propose the estimate

\[
\hat{\gamma} := \frac{4\hat{V}_f}{\sigma^2}.
\]  
(4.14)

If the values of several functions are available, then we suggest the estimate

\[
\hat{\gamma}_{ps} := \min_{1 \leq i \leq k} \frac{2\hat{V}_{f_i}}{\hat{\sigma}^2_i}.
\]  
(4.15)

Based on Proposition 1.5, the values of \( t_{mix} \) and \( 1/\gamma \) (or \( 2/\gamma_{ps} \)) are equal up to logarithmic factors. Therefore, we propose the estimate \( \hat{t}_{mix} := 1/\hat{\gamma} \) for reversible chains, and \( \hat{t}_{mix} := 2/\hat{\gamma}_{ps} \) for non-reversible chains.

5. Simulations

In the following, we present simulation results to demonstrate the applicability of the introduced error bounds. We are interested in the empirical tail probabilities of estimates, obtained from multiple runs of MCMC simulations. In particular, we will estimate logarithms of tail probabilities of the following form:

\[
\log \left( \mathbb{P} \left( \frac{\sum_{i=t_0+1}^{N} f(X_i)}{N-t_0} \geq \mathbb{E}_\pi f + t \right) \right).
\]  
(5.1)

We simulate \( m \) parallel chains and denote the sequence of states of the \( j \)th chain \( (1 \leq j \leq m) \) by \( X_1^{(j)}, \ldots, X_N^{(j)} \). Then the empirical average obtained by the \( j \)th chain can be written as

\[
\hat{E}^{(j)} := \frac{\sum_{i=t_0+1}^{N} f(X_i^{(j)})}{N-t_0},
\]  
(5.2)

and denote

\[
\hat{E} := \frac{1}{m} \sum_{j=1}^{m} \hat{E}^{(j)}.
\]  
(5.3)

Define the mean-shifted empirical distribution of these estimates as

\[
\tilde{F}(t) := \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}[\hat{E}^{(j)} - \hat{E} \leq t],
\]  
(5.4)
and let
\[ \hat{L}(t) := \begin{cases} \log \left( \hat{F}(t) \right) & \text{for } t < 0, \\ \log \left( 1 - \hat{F}(t) \right) & \text{for } t \geq 0, \end{cases} \] (5.5)
thus \( \hat{L}(t) \) is an estimate of the log tails in (5.1).

5.1. Lattice models in statistical physics

We first consider simulations on the Curie-Weiss model and the Ising model (1D and 2D). These models and their variants are widely studied in the context of MCMC simulations, and for some special cases, the mixing time and spectral gap of the chain are known.

5.1.1. Definition of models

Let \( \Omega := \{-1, 1\}^{n_s} \), and let us assume that \( \omega := (\omega_1, \ldots, \omega_{n_s}) \in \Omega \) are spins taking values 1 or -1, and distributed according to the probability distribution of the model, which is of the form (for some \( \beta > 0 \))
\[ P(\omega_1, \ldots, \omega_{n_s}) = \frac{\exp(H_{\beta,h}(\omega))}{Z}, \] (5.6)
where \( H_{\beta,h}(\omega) \) is the energy function, \( \beta \) is the inverse temperature, \( h \) corresponds to the external field, and \( Z = \sum_{\omega} \exp(H_{\beta,h}(\omega)) \) is the partition function.

In the case of the Curie-Weiss model, we define the energy function as
\[ H_{\beta,h}(\omega) = H_{\beta,h}^{CW}(\omega) := \frac{\beta}{n_s} \sum_{1 \leq i < j \leq n_s} \omega_i \omega_j + h \sum_{i=1}^{n_s} \omega_i. \] (5.7)

In the case of Ising model on a graph \( G = (V, E) \), we say that \( i \sim j \) if there is an edge between \( i \) and \( j \) in \( G \), and define \( H \) as
\[ H_{\beta,h}(\omega) = H_{\beta,h}^{I}(\omega) := \beta \sum_{i \sim j} \omega_i \omega_j + h \sum_{i=1}^{n_s} \omega_i. \] (5.8)

In the 1-dimensional Ising model, \( G \) consists of the edges \((i, i+1)\) for \( 1 \leq i \leq n_s - 1 \), while in the 2-dimensional case, \( G \) consists of the edges on a square lattice. We use periodic boundary conditions so that each spin is connected with the same number of other spins.

In practice it is impossible to obtain independent samples from (5.6). Therefore one typically designs a Markov chain which has (5.6) as its stationary distribution. In the following section we present analytic estimates of the spectral gap and mixing time of the sampling schemes defined in Section 1.1.
5.1.2. Spectral gap and mixing time

In the case of the Curie-Weiss model, at high temperature, with no external field, \((\beta < 1 \text{ and } h = 0)\), Theorem 1 of Ding, Lubetzky and Peres (2009) shows that for the Glauber dynamics, the spectral gap and mixing time are

\[
\gamma_{\text{CW}} = \frac{(1 + o(1))(1 - \beta)}{n_s}, \quad (t_{\text{mix}})_{\text{CW}} = \frac{1}{2} n_s \log \left( \frac{(1 - \beta)^2 n_s}{1 - \beta} \right) + O \left( \frac{n_s}{1 - \beta} \right),
\]

so we will use the following approximate values:

\[
\hat{\gamma}_{\text{CW}} := \frac{1 - \beta}{n_s}, \quad (\hat{t}_{\text{mix}})_{\text{CW}} := \frac{1}{2} n_s \log \left( \frac{(1 - \beta)^2 n_s}{1 - \beta} \right). \tag{5.9}
\]

Ding, Lubetzky and Peres (2009) also shows that for the critical \((\beta = 1, h = 0)\) case, the mixing time is \(O \left( n_s^{3/2} \right)\). For low temperature \((\beta > 1, h = 0)\), the mixing time is exponential in \(n_s\). The inverse spectral gap, \(1/\gamma\), has the same order as the mixing time for both the critical and the low temperature case, respectively.

A frequently used version of the Metropolis dynamics is Metropolis dynamics with spin-flip proposal (we choose a spin uniformly, and switch it to the opposite). In the case of the Curie-Weiss model, no theoretical estimates are available for the spectral gap and the mixing time of this chain, so we will estimate them from the data.

For the 1 dimensional Ising model, based on Proposition 4.13 of Paulin (2014b), we set

\[
(\hat{\gamma}_{\text{1D}}) := \frac{2}{n_s} \frac{e^{-4\beta}}{1 + e^{-4\beta}}, \quad (\hat{t}_{\text{mix}})_{\text{1D}} := \frac{n_s}{2} \left( 1 + e^{4\beta} \right) \log(4n_s),
\]

\[
(\hat{\gamma}_{\text{ps,sys}}) := \frac{8e^{-4\beta} (1 + e^{-4\beta})}{(1 + 3e^{-4\beta})^2}, \quad (\hat{t}_{\text{mix}})_{\text{ps,sys}} := \frac{1}{4} \left( 3 + e^{4\beta} \right) \log(4n_s)
\]

as estimates of the spectral gap and mixing time of the Glauber dynamics with random scan, and the pseudo spectral gap and mixing time of the Glauber dynamics with systemic scan.

5.1.3. Simulation results for total magnetization

We will consider estimating the expected value of two functions, the total magnetization, defined as

\[
m(\omega) := \sum_{i=1}^{n_s} \omega_i,
\]

and the sign of the total magnetization, defined as

\[
s(\omega) := \text{sgn} \left( \sum_{i=1}^{n_s} \omega_i \right),
\]
where \( \text{sgn}(x) := 1[x > 0] - 1[x < 0] \).

In the case of no outside field (\( h = 0 \)), we have, by symmetry, \( E_\pi m = E_\pi s = 0 \), for all of our models. For some fixed \( \beta, h \), and random initial distribution (uniformly chosen in \( \Omega \)), we run \( m \) Markov chains. The simulation results are shown in Figure 1. The plots (a-e) show empirical averages of the total magnetization \( m(\omega) \), while plot (f) shows a case when the sign of the total magnetization, \( s(\omega) \) is calculated. In all of the simulations, we have set \( t_0 \) as 30\( t_{\text{mix}} \) (or 30\( \hat{t}_{\text{mix}} \) if the mixing time was not available and had to be estimated).

5.2. Bayesian model averaging

In this section we look at simulation results on the space of directed acyclic graphs (DAGs) for Bayesian model averaging.

5.2.1. Definition of the model

DAGs are commonly used to encode the factored representation of a high-dimensional joint probability distribution. Let us consider a graph \( G = (\mathcal{X}, E) \), where \( \mathcal{X} = (X_1, X_2, \ldots, X_n) \) is a set of vertices, each representing a random variable, and \( E \) is a set of directed edges between these variables. For a node \( X_i \in \mathcal{X} \), we denote the set of its parents as \( \text{Pa}(X_i) \), where \( X_j \in \text{Pa}(X_i) \) if and only if \( (X_j, X_i) \in E \). The non-descendents of a node \( \text{Nd}(X_i) \) consist of the nodes to which there is no directed path from \( X_i \). The DAG structure entails conditional independence relations among the variables of the following form:

\[
X_i \perp \perp \text{Nd}(X_i) \mid \text{Pa}(X_i), \quad 1 \leq i \leq n \tag{5.10}
\]

With these independence assumptions, the joint distribution of the variables factors as

\[
\mathbb{P}(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid \text{Pa}(X_i)) \tag{5.11}
\]

Now, given a set of observations \( D \), we attempt to make predictions about a function \( f \) of the model structure \( G \). \( D \) consists of vectors of realizations of \( (X_1, X_2, \ldots, X_n) \). One could find a single best DAG structure with respect to \( D \), and use it to calculate the value of \( f \). Instead, we follow a Bayesian model averaging approach, where we calculate the posterior probability of each possible DAG structure and use it as a weight when making the prediction. We refer to Neapolitan (2004) for a more detailed account of this approach. The prediction \( \mathbb{E}[f(G) \mid D] \) can be expressed as a weighted average of individual predictions based on each possible DAG structure \( g \):

\[
\mathbb{E}[f(G) \mid D] = \sum_{g \in \Omega_n} f(g) \mathbb{P}(G = g \mid D), \tag{5.12}
\]
Fig 1: Simulation results for lattice models. The simulation result is plotted according to (5.5). When formulas are available for the mixing time and the spectral gap (a),(b),(d) and (e)), we use these as obtained according to Section 5.1.2. Otherwise we used estimated values of the parameters \( \gamma \), \( \tilde{t}_{\text{mix}} \), \( \tilde{\sigma}^2 \), and \( \tilde{V}_f \) (see Section 4). The reversible Chebyshev and Bernstein bounds are plotted according to (3.7), and (3.8) respectively, except for the non-reversible case (e), where we used the bounds (3.10) and (3.11). We also show the quantiles of \( N(0, \tilde{\sigma}^2) \) arising from the CLT (see Section 2).
where $\Omega_n$ is the set of all DAG structures on $n$ variables.

The model $G$ is parametrized using a set of conditional probability tables describing the probability of each node taking a certain value given its parents. We denote the set of all such parameters $\theta_G$.

The posterior probability of a structure $G$ can be obtained by applying Bayes theorem on its marginal likelihood. The marginal likelihood is generally expressed as

$$P(D|G) = \int_{\theta_G} P(D|\theta_G, G)P(\theta_G)d\theta_G. \quad (5.13)$$

In our current example, we assume that each random variable is binary, that is, $X_i \in \{0, 1\}$. As is typically done in the context of binary DAG models, we set a beta distribution as the prior distribution of each variable conditioned on its parent configuration. Using beta priors, Heckerman, Geiger and Chickering (1995) shows that the marginal likelihood can be calculated as

$$P(D|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(s_{ij})}{\Gamma(d_{ij} + s_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(d_{ijk} + s_{ijk})}{\Gamma(s_{ijk})}, \quad (5.14)$$

where $i$ refers to a node $X_i$, $j$ is a value configuration of the parents of node $X_i$, with $q_i$ the total number of parent value configurations, $k$ indicates the value of node $X_i$ under parent configuration $j$, and $r_i$ is the number of different values that $X_i$ can take. For each combination of indices, $d_{ij}$ and $d_{ijk}$ represent the observed count, while $s_{ij}$ and $s_{ijk}$ are the prior counts. To make priors consistent among different DAG structures, we choose a fixed equivalent sample size $S$, and set $s_{ijk} = S/(q_ir_i)$.

For simplicity, we assume that the prior probability for each structure is equal, that is, $\forall G \in \Omega_n, P(G) = 1/|\Omega_n|$. As the number of summation terms in (5.12) can be prohibitively large to compute exactly, we design a Markov chain with stationary distribution $P(G|D)$ and use a Monte Carlo estimate to approximate the prediction.

### 5.2.2. Procedure

We follow Madigan, York and Allard (1995), and design a Markov chain on $\Omega_n$ with stationary distribution $P(G|D)$. Starting with an initial DAG structure, the chain either adds or removes a single edge at each proposal step. We denote the neighborhood of a state $G_i$ as $Nb(G_i)$, which is the set of DAGs that differ from $G_i$ by one edge addition or one edge removal. The chain then uses the following probabilities to propose the next state:

$$T(G_j|G_i) = \begin{cases} \frac{1}{|Nb(G_i)|}, & G_j \in Nb(G_i) \\ 0, & G_j \notin Nb(G_i) \end{cases}. \quad (5.15)$$

When making the proposal, we make sure that only valid (cycle-free) DAGs are considered. The chain moves to the proposed state with the following ac-
ceptance probability:
\[
A(G_j|G_i) = \min \left\{ 1, \frac{|Nb(G_i)| \mathbb{P}(G_j|D)}{|Nb(G_j)| \mathbb{P}(G_i|D)} \right\}.
\] (5.16)

Note that the ratio of marginal likelihoods can be evaluated locally at the target node of the single edge that is changed during the proposal step. As opposed to some of the lattice models discussed in the previous section, here, no analytic formulas are known for the mixing time and spectral gap of the Markov chain.

5.2.3. Simulation results

In the following simulation example, we have a set of \( n = 6 \) variables, thus the space of the Markov Chain consist of DAGs with 6 vertices. We take a data set \( D \) consisting of 20 vectors generated from a known DAG on 6 nodes (structure not shown), and assume a prior equivalent sample size of 4. Our goal is to estimate the posterior probability of an edge being present in the structure:
\[
f(G) = \begin{cases} 1, & (X_i, X_j) \in E_G \\ 0, & (X_i, X_j) \notin E_G \end{cases}.
\] (5.17)

We look at two cases, first, at the presence of the edge \( e_a = (i = 1, j = 2) \), and then at \( e_b = (i = 1, j = 4) \). The simulation results are shown in Figure 2 (a) and (b) respectively. These figures show examples of exponential tails, for which our proposed Bernstein bound provides a tight upper bound. The normal quantile based estimate is poor on the side with exponential tail.

Final remarks

In this paper, we demonstrated the practical usability of concentration inequalities for obtaining non-asymptotically valid error bounds for MCMC empirical averages. We stated Chebyshev and Bernstein-type inequalities for reversible, and non-reversible chains, in a form that is directly applicable in practice. We then proposed estimators for the all the parameters arising in the bounds, and gave theoretical explanation for their usage. We have included several simulation examples, which show the advantage of the non-asymptotic approach compared to the asymptotic approach using normal approximation.

Finally, we mention that in the case of independent random variables, rather precise error estimates with small constants are available for the normal approximation of empirical averages (see, for example Chen, Goldstein and Shao (2011)). It would be interesting to obtain similarly sharp Berry-Esseen-type results for Markov chains as well.

Acknowledgements

The second author thanks Doma Szász and Mogyi Tóth for infecting him with their enthusiasm of probability. He also thanks his brother, Roland Paulin, for
Fig 2: Simulation results for Bayesian model averaging. The simulation result is plotted according to (5.5). We use estimated values of the parameters $\hat{\gamma}$, $\hat{t}_{\text{mix}}$, $\hat{\sigma}^2$ and $\hat{V}_f$ (see Section 4), and plot the Bernstein bound according to (3.8), as well as the Chebyshev bound according to (3.7). We also show the quantiles of $N(0, \hat{\sigma}^2)$, arising from the CLT (see Section 2). For both cases we estimate $\hat{\gamma}$, and then use the minimum of the two values in the bounds in accordance with (4.13).
the enlightening discussions. The authors thank their thesis supervisors, Louis Chen, Adrian Röllin and David Hsu for the opportunity to study in Singapore, and their useful advices. We also thank Daniel Rudolf for his comments. Finally, we thank Lee Hwee Kuan for his contribution to the simulation code.

References

Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. Electron. J. Probab. 13 no. 34, 1000–1034. . MR2424985 (2009j:60032)

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

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Statist. 7 434–455. . MR1665662 (99k:62055)

Brooks, S. P. and Roberts, G. O. (1998). Assessing convergence of Markov chain Monte Carlo algorithms. Statistics and Computing 8 319–335.

Chen, L. H. Y., Goldstein, L. and Shao, Q.-M. (2011). Normal approximation by Stein’s method. Probability and its Applications (New York). Springer, Heidelberg. MR2732624

Chen, F., Lovász, L. and Pak, I. (1999). Lifting Markov chains to speed up mixing. In Annual ACM Symposium on Theory of Computing (Atlanta, GA, 1999) 275–281 (electronic). ACM, New York. . MR1798046 (2001i:68178)

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Amer. Statist. Assoc. 91 883–904. . MR1395755

Diaconis, P., Holmes, S. and Neal, R. M. (2000). Analysis of a nonreversible Markov chain sampler. Ann. Appl. Probab. 10 726–752. . MR1789978 (2001i:60114)

Ding, J., Lubetzky, E. and Peres, Y. (2009). The mixing time evolution of Glauber dynamics for the mean-field Ising model. Comm. Math. Phys. 289 725–764. . MR2506768 (2010e:82064)

Douc, R., Moulines, E., Olsson, J. and van Handel, R. (2011). Consistency of the maximum likelihood estimator for general hidden Markov models. Ann. Statist. 39 474–513. . MR2797854 (2012b:62277)

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical science 7 457–472.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). Bayesian data analysis, second ed. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL. MR2027492 (2004j:62001)

Geyer, C. J. (1992a). Markov chain Monte Carlo maximum likelihood. Defense Technical Information Center.

Geyer, C. J. (1992b). Practical Markov chain Monte Carlo. Statistical Science 7 473–483.
Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., eds. (1996). Markov chain Monte Carlo in practice. Interdisciplinary Statistics. Chapman & Hall, London. MR1397966 (97d:62006)

Gillman, D. (1998). A Chernoff bound for random walks on expander graphs. *SIAM J. Comput.* 27 1203–1220. MR1621958 (99e:60076)

Glynn, P. W. and Ormoneit, D. (2002). Hoeffding’s inequality for uniformly ergodic Markov chains. *Statist. Probab. Lett.* 56 143–146. MR1881167 (2002k:60143)

Heckerman, D., Geiger, D. and Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20 197–243.

Jarner, S. F. and Roberts, G. O. (2002). Polynomial convergence rates of Markov chains. *Ann. Appl. Probab.* 12 224–247. MR1890063 (2003c:60117)

Jarner, S. F. and Tweedie, R. L. (2003). Necessary conditions for geometric and polynomial ergodicity of random-walk-type Markov chains. *Bernoulli* 9 559–578. MR1996270 (2003c:60117)

Joulin, A. and Ollivier, Y. (2010). Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.* 38 2418–2442. MR2683634 (2011j:60229)

Latuszynski, K., Miasojedow, B. and Niemiro, W. (2011). Nonasymptotic bounds on the estimation error of MCMC algorithms. arXiv preprint arXiv:1106.4739.

León, C. A. and Perron, F. (2004). Optimal Hoeffding bounds for discrete reversible Markov chains. *Ann. Appl. Probab.* 14 958–970. MR2052909 (2005d:60109)

Levin, D. A., Peres, Y. and Wilmer, E. L. (2009). Markov chains and mixing times. American Mathematical Society, Providence, RI. With a chapter by James G. Propp and David B. Wilson. MR2466937 (2010c:600209)

Lezaud, P. (1998a). Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.* 8 489–867. MR1627795 (99f:60061)

Lezaud, P. (1998b). Etude quantitative des chaînes de Markov par perturbation de leur noyau. Thèse doctorat mathématiques appliquées de l’Université Paul Sabatier de Toulouse, Available at http://pom.tls.cena.fr/papers/thesis/these_lezaud.pdf.

Lezaud, P. (2001). Chernoff and Berry-Essèen inequalities for Markov processes. *ESAIM Probab. Statist.* 5 183–201. MR1875670 (2003c:60168)

Lindvall, T. (1992). Lectures on the coupling method. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication. MR1180522 (94c:60002)

Liu, J. S. (2008). Monte Carlo strategies in scientific computing. *Springer Series in Statistics*. Springer, New York. MR2401592 (2010b:65013)

Madigan, D., York, J. and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique* 215–232.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos
Neapolitan, R. E. (2004). *Learning Bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ.

Ollivier, Y. (2009). Ricci curvature of Markov chains on metric spaces. *J. Funct. Anal.* **256**, 810–864. MR2484937 (2010j:58081)

Paulin, D. (2014a). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *arXiv preprint*.

Paulin, D. (2014b). Mixing and Concentration by Ricci Curvature. *arXiv preprint*.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*, second ed. Springer Texts in Statistics. Springer-Verlag, New York. MR2080278 (2005d:62006)

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.* **1**, 20–71. MR2095565 (2005i:60135)

Rubinstein, R. Y. and Kroese, D. P. (2008). *Simulation and the Monte Carlo method*, second ed. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. MR2365210 (2008i:68002)

Rudolf, D. (2011). Explicit error bounds for Markov chain Monte Carlo. *arXiv preprint arXiv:1108.3201*.

Samson, P.-M. (2000). Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.* **28**, 416–461. MR1756011 (2001d:60015)

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*. 
Appendix

In this section, we will prove our propositions in Section 4, bounding the error of estimators \( \hat{V}_f \) and \( \hat{\sigma}^2(k) \).

**Proposition 5.1.** Suppose that \( X_1, \ldots, X_N \) is an uniformly ergodic Markov chain, with stationary distribution \( \pi \), and initial distribution \( q \). For any \( T \geq 0 \),

\[
\mathbb{P}_q \left( \frac{V_f - \hat{V}_f}{N - t_0} + T \geq \frac{-(N - \hat{t}_0)}{8C^4t_{\text{mix}}} + E(\hat{t}_0) \right) = \exp \left( -\frac{T^2(N - \hat{t}_0)}{200t_{\text{mix}}C^4} \right).
\]

(5.18)

**Proof of Proposition 5.1.** Changing the function \( f \) to \( f - \mathbb{E}_\pi f \) does not changes \( \hat{V}_f \), so we can assume that \( \mathbb{E}_\pi f = 0 \), and \( |f(x)| \leq C \). Now it is easy to show that \( \hat{V}_f \) changes at most by \( 5C^2/(N - \hat{t}_0) \) is we change \( X_i \). From this, using McDiarmid’s bounded differences inequality for Markov chains (Corollary 2.9 of Paulin (2014a)), we can deduce that for any \( T \geq 0 \),

\[
\mathbb{P}_\pi (\hat{V}_f - \mathbb{E}_\pi \hat{V}_f \geq T) \leq \exp \left( -\frac{T^2(N - \hat{t}_0)}{200t_{\text{mix}}C^4} \right).
\]

Moreover, we have

\[
V_f - \mathbb{E}_\pi \hat{V}_f = \text{Var}_\pi \left( \frac{1}{N - t_0} \sum_{i=t_0+1}^N f(X_i) \right),
\]

which, by Theorems 3.5 and 3.7 of Paulin (2014a), can be further bounded by \( 2V_f/\gamma/(N - \hat{t}_0) \) for reversible chains, and by \( 4V_f/\gamma_{ps}/(N - \hat{t}_0) \) for non-reversible chains. Using Proposition 1.5, these can be further bounded by \( 8t_{\text{mix}}/(N - \hat{t}_0) \), and the result follows.

We will use the following lemma for the proof of our propositions about \( \hat{\sigma}^2(k) \).

**Lemma 5.2.** 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,

\[
- \min \left( \frac{V_f}{2}, \frac{2V_f}{\gamma} \right) (1 - \gamma)^{k+1} \leq \sigma^2 - \left( \gamma_0 + 2 \sum_{t=1}^k \gamma_t \right) \leq \frac{2V_f}{\gamma} (1 - \min(\gamma, 1))^{k+1}.
\]

(5.19)

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

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

(5.20)
Proof. Without loss of generality, assume that $\mathbb{E}_\pi f = 0$. Define the operator $\pi$ on $L^2(\pi)$ as $\pi(g)(x) := \mathbb{E}_\pi(g)$. We have $\sigma^2 = \gamma_0 + 2 \sum_{i=1}^{\infty} \gamma_i$, thus

$$
\sigma^2 - \left( \gamma_0 + 2 \sum_{i=1}^{k} \gamma_i \right) = 2 \sum_{i=k+1}^{\infty} \gamma_i = 2 \left\langle f, \left( \sum_{i=k+1}^{\infty} P_i \right) f \right\rangle_{\pi}
$$

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

For reversible chains, we have $\|P - \pi\|_{2,\pi} \leq 1 - \gamma^*$, and $\|(I - (P - \pi))^{-1}\|_{2,\pi} = 1/\gamma$, thus

$$
\left| \sigma^2 - \left( \gamma_0 + 2 \sum_{i=1}^{k} \gamma_i \right) \right| \leq \frac{2V_f}{\gamma} \cdot (1 - \gamma^*)^{k+1}. \quad (5.21)
$$

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

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

Now it is easy to see that

$$
\| (P - \pi)^{k+1}_+ (I - (P - \pi))^{-1} \|_{2,\pi} \leq \min(\gamma, 1)^{k+1}/\gamma, \quad \text{and}
$$

$$
\| (P - \pi)^{k+1}_- (I - (P - \pi))^{-1} \|_{2,\pi} \leq 1/2,
$$

thus

$$
- \min \left( V_f, \frac{2V_f}{\gamma} (1 - \gamma^*)^{k+1} \right) \leq 2 \left\langle f, (P - \pi)^{k+1}_+ (I - (P - \pi))^{-1} f \right\rangle_{\pi}
$$

$$
\leq \frac{2V_f}{\gamma} (1 - \min(\gamma, 1))^{k+1}.
$$

Combining this and (5.21) leads to (5.19).

For non-reversible chains, by the proof of Theorem 3.7 of Paulin (2014a), we have that $\| (I - (P - \pi))^{-1} \|_{2,\pi} \leq 2/\gamma_{ps}$, and $\|(P - \pi)^{k+1} \|_{2,\pi} \leq (1 - \gamma_{ps})^{(k+1-1)/2}$, thus (5.20) follows.

First, we prove the bounds on the bias of $\hat{\sigma}^2(k)$, and then the concentration inequality.

Proof of Proposition 4.2. For reversible chains, for $0 \leq i \leq k$, from Chebyshev’s inequality (Theorem 3.2), we get

$$
\left| \mathbb{E}_{\pi} \left( \frac{1}{2} \left( \frac{\sum_{j=t_0+1}^{N-k} f(X_j)}{N - t_0 - k} \right)^2 + \frac{1}{2} \left( \frac{\sum_{j=t_0+1}^{N-k+1} f(X_j)}{N - t_0 - k} \right)^2 \right) - \left( \mathbb{E}_f f \right)^2 \frac{\sigma^2}{N - t_0 - k} \right|
$$

$$
\leq \frac{4V_f}{\gamma^2} \cdot \frac{1}{(N - t_0 - k)^2}.
$$
and thus it follows that
\[
\left| \mathbb{E}_\pi (\hat{\gamma}_i) - \left( \gamma_i - \frac{\sigma^2}{N - t_0 - k} \right) \right| \leq \frac{4V_f}{\gamma^2} \cdot \frac{1}{(N - t_0 - k)^2}.
\]
Summing up in \( i \), and using (5.19) leads to
\[
- K_f - \min \left( V_f, \frac{2V_f}{\gamma} (1 - \gamma^*)^{k+1} \right) \leq \sigma^2 - \left( \hat{\gamma}_0 + 2 \sum_{i=1}^k \hat{\gamma}_i + \frac{\sigma^2 (2k + 1)}{N - t_0 - k} \right)
\]
\[
\leq K_f + \frac{2V_f}{\gamma} \cdot (1 - \min(\gamma, 1))^{k+1},
\]
where \( K_f := \frac{4V_f}{\gamma^2} \cdot \frac{(2k+1)}{(N - t_0 - k)^2} \). Now putting together the terms involving \( \sigma^2 \), and dividing by \( \frac{N - t_0 - 3k - 1}{N - t_0 - k} \) leads to (4.5). The proof of (4.6) is similar.

Proof of Proposition 4.3. 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.

A simple computation shows that changing the value of \( X_j \), for \( t_0 + 1 \leq j \leq \hat{N} \), can only change \( \hat{\gamma}_i \) at most by \( 8C^2/(\hat{N} - t_0 - k) \), and thus it can only change the value of \( \hat{\sigma}^2(k) \) at most by \( 8(2k + 1)C^2/(\hat{N} - t_0 - 3k - 1) \). From this (the so called Hamming-Lipschitz property), using McDiarmid’s bounded differences inequality for Markov chains (Corollary 2.9 of Paulin (2014a)), we can deduce (4.7). Finally, (4.8) and (4.9) follow by combining this with the bounds on the bias.