Error Bounds for Computing the Expectation by Markov Chain Monte Carlo

D. Rudolf

Preprint 21
Error Bounds for Computing the Expectation by Markov Chain Monte Carlo

D. Rudolf

Preprint 21
The consecutive numbering of the publications is determined by their chronological order.

The aim of this preprint series is to make new research rapidly available for scientific discussion. Therefore, the responsibility for the contents is solely due to the authors. The publications will be distributed by the authors.
ERROR BOUNDS FOR COMPUTING THE
EXPECTATION BY MARKOV CHAIN MONTE CARLO

DANIEL RUDOLF

ABSTRACT. We study the error of reversible Markov chain Monte Carlo methods for approximating the expectation of a function. Explicit error bounds with respect to the \( l_2 \), \( l_4 \) and \( l_\infty \)-norm of the function are proven. By the estimation the well known asymptotical limit of the error is attained, i.e. there is no gap between the estimate and the asymptotical behavior. We discuss the dependence of the error on a burn-in of the Markov chain. Furthermore we suggest and justify a specific burn-in for optimizing the algorithm.

1. INTRODUCTION

We start with a probability distribution \( \pi \) on a finite set \( D \) and a function \( f : D \to \mathbb{R} \). The goal is to compute the expectation denoted by

\[
S(f) = \sum_{x \in D} f(x) \pi(x).
\]

Let the cardinality of \( D \) be very large such that an exact computation of the sum is practically impossible. Furthermore suppose that the desired distribution is not explicitly given, i.e. we have no random number generator for \( \pi \) available. Such kind of problems arise in statistical physics, in statistics, and in financial mathematics (see for instance [GRS96, Liu08]). The idea of approximating \( S(f) \) via Markov chain Monte Carlo (MCMC) is the following: Run a Markov chain on \( D \) to simulate the distribution \( \pi \) and compute the time average over the last \( n \) steps. Let \( X_1, \ldots, X_{n+n_0} \) be the chain, then we obtain as approximation

\[
S_{n,n_0}(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_{i+n_0}).
\]

---

Date: Version: July 28, 2009.

Key words and phrases. Markov chain Monte Carlo methods, Markov chain Monte Carlo, error bounds, explicit error bounds, burn-in, mixing time, eigenvalue.
By \( n_0 \) the so called burn-in is given, loosely spoken this is the number of time steps taken to warm up. Afterwards the distribution of the generated Markov chain is (hopefully) close to the stationary one.

A Markov chain is identified with its initial distribution \( \nu \) and its transition matrix \( P \). We restrict ourself to ergodic chains, i.e. the second largest absolute value \( \beta \) of the eigenvalues of \( P \) is smaller than one. It is well known that the distribution of these chains reaches stationarity exponentially (see [Bre99, RR97, LPW09]).

The error of \( S_{n,n_0} \) for \( f \in \mathbb{R}^D \) is measured by

\[
e_{\nu}(S_{n,n_0}, f) = \left( E_{\nu,P} \left| S_{n,n_0}(f) - S(f) \right|^2 \right)^{1/2},
\]

where \( E_{\nu,P} \) denotes the expectation of the Markov chain. The asymptotic behavior of the integration error can be written in terms of the eigenvalues and eigenfunctions of \( P \). It holds true that

\[
\lim_{n \to \infty} n \cdot e_{\nu}(S_{n,n_0}, f)^2 \leq \frac{1 + \beta_1}{1 - \beta_1} \| f \|_2^2,
\]

where \( \beta_1 \) is the second largest eigenvalue (see [Sok97, Mat99]). The constant \( \frac{1 + \beta_1}{1 - \beta_1} \) is optimal but this statement does not give an error bound for finite \( n \) and also does not include anything concerning the choice of \( n_0 \). How does an explicit error bound of the MCMC method look like where the asymptotic behavior is attained?

Let us give an outline of the structure and the main results. Section \( 2 \) contains the used notation and presents some relevant statements concerning Markov chains. Section \( 3 \) contains the new results. The explicit error bound is developed with respect to the \( l_2 \), \( l_4 \) and \( l_{\infty} \)-norm of the function \( f \). For \( \| f \|_\infty \leq 1 \) and \( C = 2 \sqrt{\| \frac{\nu}{\pi} - 1 \|_\infty} \) we obtain the following. The error obeys

\[
e_{\nu}(S_{n,n_0}, f)^2 \leq \frac{2}{n(1 - \beta_1)} + \frac{2C\beta^{n_0}}{n^2(1 - \beta)^2}.
\]

For details and estimates concerning \( l_2 \) and \( l_4 \) we refer to Theorem 11 in Section 3.3. In Section 4 it turns out that \( n_0 = \max \left\{ \left\lfloor \frac{\log(C)}{\log(\beta - 1)} \right\rfloor, 0 \right\} \) is a reasonable choice for the burn-in. Then the error bound simplifies to

\[
e_{\nu}(S_{n,n_0}, f)^2 \leq \frac{2}{n(1 - \beta_1)} + \frac{2}{n^2(1 - \beta)^2}.
\]
In many examples a good estimate for $\beta$ can be achieved, see for instance [MR02, BD06, BL07]. Therefore it is straightforward to apply the explicit error bound.

2. Preliminaries

The Markov chain $X_1, X_2, \ldots$ is a stochastic process with state space $D$. We identify it with initial distribution $\nu$ and transition matrix $P = (p(x, y))_{x, y \in D}$ and denote it by $(\nu, P)$. For $x, y \in D$ the entry $p(x, y)$ presents the probability of jumping from state $x$ to state $y$ in one step of the chain.

By $Pf(x) = \sum_{y \in D} p(x, y)f(y)$ we obtain the expectation of the value of $f \in \mathbb{R}^D$ after one step of the chain starting from $x \in D$. The expectation after $k$ steps of the Markov chain from $x$ is given by $P^k f(x) = \sum_{y \in D} p^k(x, y)f(y)$, where $P^k = (p^k(x, y))_{x, y \in D}$ denotes the $k$-th power of $P$. Similarly we consider the application of $P$ to a distribution $\nu$, i.e. $\nu P(x) = \sum_{y \in D} \nu(y)p(y, x)$. This is the distribution which arises after one step where the initial state was chosen by $\nu$. The distribution which arises after $k$ steps is given by $\nu P^k(x) = \sum_{y \in D} \nu(y)p^k(y, x)$.

The expectation $E_{\nu, P}$ of the Markov chain $X_1, \ldots, X_{n+n_0}$ is taken with respect to the probability measure

$$W_{\nu, P}(x_1, \ldots, x_{n+n_0}) = \nu(x_1)p(x_1, x_2) \cdots p(x_{n+n_0-1}, x_{n+n_0})$$
on $D^{n+n_0}$. Using this for $i \leq j$ we obtain a characterization by the transition matrix

$$E_{\nu, P}(f(X_i)f(X_j)) = \sum_{x \in D} P^i(fP^{j-i}f)(x)\nu(x).$$

2.1. Reversibility and spectral structure. We call the Markov chain with transition matrix $P$, or simply $P$, reversible with respect to a probability measure $\pi$ if the detailed balance condition

$$\pi(x)p(x, y) = \pi(y)p(y, x)$$

holds true for $x, y \in D$. If $P$ is reversible, then $\pi$ is a so called stationary distribution of the Markov chain, i.e. $\pi P(x) = \pi(x)$. Note that, if $P$ is reversible then $P^k$ is also reversible. Let us define the weighted scalar-product

$$\langle f, g \rangle_\pi = \sum_{x \in D} f(x)g(x)\pi(x),$$
for functions $f, g \in \mathbb{R}^D$. Then let $\|f\|_2 = \langle f, f \rangle_\pi^{1/2}$. By considering the scalar-product it is easy to show, that reversibility is equivalent to $P$ being self-adjoint. Furthermore suppose that the underlying Markov chain is irreducible and aperiodic, this is also called ergodic. For details of these conditions we refer to the literature, for instance [Häg02, Bré99, LPW09]. It is a well known fact that this implies the uniqueness of the stationary distribution. Applying the spectral theorem of self-adjoint stochastic matrices and ergodicity we obtain that $P$ has real eigenvalues

$$1 = \beta_0 > \beta_1 \geq \beta_2 \geq \cdots \geq \beta_{|D|-1} > -1$$

with a basis of orthogonal eigenfunctions $u_i$ for $i \in \{0, \ldots, |D| - 1\}$, i.e.

$$Pu_i = \beta_i u_i, \quad \langle u_i, u_j \rangle_\pi = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

Additionally one can see that $u_0(x) = 1$ and $S(u_i) = 0$ for $i > 0$.

### 2.2. Convergence of the chain.

The speed of convergence of the Markov chain to stationarity is measured by the so called $\chi^2$-contrast. Let $\nu, \mu$ be distributions on $D$ then

$$\chi^2(\nu, \mu) = \sum_{x \in D} \frac{(\nu(x) - \mu(x))^2}{\mu(x)}.$$ 

The $\chi^2$-contrast is not symmetric and therefore no distance. For arbitrary distributions it can be very large, i.e. $\chi^2(\nu, \mu) \leq \left\| \frac{\nu}{\mu} - 1 \right\|_\infty$, where

$$\left\| \frac{\nu}{\mu} - 1 \right\|_\infty = \max_{x \in D} \left| \frac{\nu(x)}{\mu(x)} - 1 \right|.$$ 

From [Bré99 Theorem 3.3 p. 209] we have

$$(2) \quad \chi^2(\nu P^k, \pi) \leq \beta^k \chi^2(\nu, \pi),$$
where \( \beta = \max \{ \beta_1, |\beta_{|D|-1}| \} \) denotes the second largest absolute value of the eigenvalues. Let us turn to another presentation of the convergence property. We have

\[
\nu P^k(x) - \pi(x) = \sum_{y \in D} \frac{\nu(y)}{\pi(y)} p^k(y, x) \pi(y) - \pi(x)
\]

\[
= \sum_{y \in D} \nu(y) p^k(x, y) \pi(x) - \pi(x)
\]

\[
= \sum_{y \in D} \frac{\nu(y)}{\pi(y)} p^k(x, y) \pi(x) - \sum_{y \in D} \frac{\nu(y)}{\pi(y)} \pi(y) \pi(x)
\]

\[
= \sum_{y \in D} \frac{\nu(y)}{\pi(y)} (p^k(x, y) - \pi(y)) \pi(x).
\]

The second equality follows by the reversibility of the Markov chain.

For simplicity let

\[
d_k(x) := \sum_{y \in D} \frac{\nu(y)}{\pi(y)} (p^k(x, y) - \pi(y)),
\]

such that altogether

\[
\|d_k\|_2 = \sqrt{\chi^2(\nu P^k, \pi)} \leq \beta \sqrt{\left\| \frac{\nu}{\pi} - 1 \right\|_\infty}.
\]

Since \( \beta < 1 \) we have an exponential decay of the norm with \( k \to \infty \).

We define the weighted sequence spaces for \( 1 \leq p \leq \infty \) by

\[
l_p = l_p(D, \pi) := \left\{ f \in \mathbb{R}^D : \|f\|_p = \sum_{x \in D} |f(x)|^p \pi(x) < \infty \right\}.
\]

It is clear that \( l_p = \mathbb{R}^D \), since the state space has finite cardinality.

**Remark 1.** As we have seen the \( \chi^2 \)-contrast corresponds to the \( l_2 \)-norm of the function \( d_k \). Other tools for measuring the speed of convergence induce similar relations. For instance

\[
\|d_k\|_1 = 2 \|\nu P^k - \pi\|_{1v} \quad \text{and} \quad \|d_k\|_\infty = \left\| \frac{\nu P^k}{\pi} - 1 \right\|_{\infty}.
\]

The total variation corresponds to the \( l_1 \)-norm of \( d_k \) and the \( l_\infty \)-norm to the supremum-distance.

**Remark 2.** The constant \( \beta \) plays a crucial role in estimating the speed of convergence of the Markov chain to stationarity. In general it is not easy to handle \( \beta_1 \) or \( \beta \), but there are different auxiliary tools, e.g.
canonical path technique, conductance (see [JS89] and [DS91]), log-
Sobolev inequalities and path coupling. For a small survey see [Ran06].

2.3. Norm of the transition matrix. Let us consider $P$ and $S$ as operators acting on $l_p$. Then the functional $S$ maps arbitrary functions to constant functions. Let

$$l^0_p := l^0_p(D, \pi) = \{g \in l_p : S(g) = 0\} \quad \text{for } 2 \leq p \leq \infty.$$ 

The norm of $P$ as operator on $l^0_2$ and $l^0_4$ is essential in the analysis. We state and show some results which are implied by the Theorem of Riesz-Thorin. For a proof and an introduction we refer to [BS88].

**Proposition 1** (Theorem of Riesz-Thorin). Let $1 \leq p, q_1, q_2 \leq \infty$. Further let $\theta \in (0, 1)$ and

$$\frac{1}{p} := \frac{1 - \theta}{q_1} + \frac{\theta}{q_2}$$

and

$$T : l_{q_1} \rightarrow l_{q_1} \quad \text{with} \quad \|T\|_{l_{q_1} \rightarrow l_{q_1}} \leq M_1,$$

$$T : l_{q_2} \rightarrow l_{q_2} \quad \text{with} \quad \|T\|_{l_{q_2} \rightarrow l_{q_2}} \leq M_2.$$ 

Then

$$\|T\|_{l_p \rightarrow l_p} \leq 2M_1^{1-\theta}M_2^\theta.$$ 

Note that the factor two in the last inequality comes from the fact that we consider real valued functions $f$. In the following we show a relation between $P$, $P - S$ and $\beta$.

**Lemma 2.** Let $P$ be a reversible transition matrix with respect to $\pi$ and $n \in \mathbb{N}$. Then

$$\|P^n - S\|_{l_2 \rightarrow l_2} = \|P^n\|_{l^0_2 \rightarrow l^0_2} = \beta^n.$$ 

Furthermore if $2 \leq p \leq \infty$ then

$$\|P^n\|_{l^p \rightarrow l^p} \leq \|P^n - S\|_{l_p \rightarrow l_p} \leq 2.$$ 

**Proof.** The self-adjointness of $P$ implies $\|P\|_{l^q_2 \rightarrow l^q_2} = \max \{\beta_1, |\beta_{|E|-1}|\} = \beta$, such that $\|P^n\|_{l^q_2 \rightarrow l^q_2} = \beta^n$. By

$$\|P^n - S\|_{l_2 \rightarrow l_2} = \sup_{\|f\|_2 \leq 1} \|(P^n - S)f\|_2 = \sup_{\|f\|_2 \leq 1} \|P^n(f - S(f))\|_2 \leq \sup_{\|f\|_2 \leq 1} \|P^n g\|_2 = \|P^n\|_{l^0_2 \rightarrow l^0_2}$$

and

$$\|P^n - S\|_{l_p \rightarrow l_p} \leq \|P^n\|_{l^p \rightarrow l^p}.$$
and
\[ \|P^n\|_{l_0 \rightarrow l_0} = \sup_{\|g\|_p \leq 1, \|S\|_p = 0} \|P^n g - S(g)\|_p \leq \sup_{\|f\|_p \leq 1} \|(P^n - S)f\|_p = \|P^n - S\|_{l_p \rightarrow l_p} \]

claim (4) and the first part of (5) is shown. Finally, by applying the triangle inequality of the norm
\[ \|P^n - S\|_{l_p \rightarrow l_p} = \|P^n\|_{l_p \rightarrow l_p} + \|S\|_{l_p \rightarrow l_p} = 2. \]

□

3. Error bounds

In this section we mainly follow two steps to develop the error bound. At first a special case of method \( S_{n,n_0} \) is considered. The initial distribution is the stationary one, thus it is not necessary to do a burn-in, i.e. \( n_0 = 0 \). Secondly we relate the result of the first step to the general case where the chain is initialized by a distribution \( \nu \). The techniques which we will use are similar as in [Rud09].

3.1. Starting from stationarity. This is also called starting in equilibrium, i.e. the distribution of the Markov chain does not change, it is already balanced. In the following we will always denote \( S_{n,0} \) as \( S_n \). Let us start with stating and discussing a result from [BD06, Prop. 2.1 p.3].

**Proposition 4.** Let \( f \in \mathbb{R}^D \). Let \( X_1, \ldots, X_n \) be a reversible Markov chain with respect to \( \pi \), given by \( (P, \pi) \). Then
\[ e_\pi(S_n, f)^2 = \frac{1}{n^2} \sum_{k=1}^{\lvert D \rvert - 1} |a_k|^2 W(n, \beta_k), \]

where \( a_k = \langle f, u_k \rangle_\pi \) and \( W(n, \beta_k) := \frac{n(1 - \beta_k^2) - 2\beta_k(1 - \beta_k^n)}{(1 - \beta_k^2)^2} \).

**Proof.** Let us consider \( g := f - S(f) \in \mathbb{R}^D \). Because of the orthogonal basis the presentation \( g(x) = \sum_{k=1}^{\lvert D \rvert - 1} a_k u_k(x) \) is given. The error obeys

\[
e(S_n, f)^2 = E_{\pi, P} \left| \frac{1}{n} \sum_{j=1}^{n} g(X_j) \right|^2 = \frac{1}{n^2} E_{\pi, P} \left| \sum_{j=1}^{n} g(X_j) \right|^2
\]

\[
= \frac{1}{n^2} \sum_{j=1}^{n} E_{\pi, P} g(X_j)^2 + \frac{2}{n^2} \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} E_{\pi, P} g(X_j) g(X_i).
\]

For \( i \leq j, \)

\[
E_{\pi, P} g(X_i) g(X_j) = \sum_{k=1}^{\lvert D \rvert - 1} a_k a_l E_{\pi, P} u_k(X_i) u_l(X_j)
\]

\[
= \sum_{k=1}^{\lvert D \rvert - 1} a_k a_l \langle u_k, P^{j-i} u_l \rangle_\pi
\]

\[
= \sum_{k=1}^{\lvert D \rvert - 1} a_k a_l \beta_{j-i} \langle u_k, u_l \rangle_\pi = \sum_{k=1}^{\lvert D \rvert - 1} a_k^2 \beta_{j-i}^{j-i},
\]

where the equality of the second line is due to the fact that the initial step is chosen from the stationary distribution. The last two equalities follow from the orthonormality of the basis of the eigenvectors. Altogether we have

\[
e(S_n, f)^2 = \frac{1}{n^2} \sum_{k=1}^{\lvert D \rvert - 1} a_k^2 \left[ n + 2 \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} \beta_k^{i-j} \right]
\]

\[
= \frac{1}{n^2} \sum_{k=1}^{\lvert D \rvert - 1} a_k^2 \left[ n + 2 \frac{(n-1)\beta_k - n\beta_k^2 + \beta_k^{n+1}}{(1 - \beta_k)^2} \right]
\]

\[
= \frac{1}{n^2} \sum_{k=1}^{\lvert D \rvert - 1} |a_k|^2 W(n, \beta_k).
\]
Let us consider $W(n, \beta_k)$ to simplify and interpret Proposition 4.

**Lemma 5.** For all $n \in \mathbb{N}$ and $k \in \{1, \ldots, |D| - 1\}$ we have

$$W(n, \beta_k) \leq W(n, \beta_1) \leq \frac{2n}{1 - \beta_1}.$$  

*Proof.* Let $x \in [-1, 1)$, then we are going to show that $W(n, x)$ is monotone increasing, i.e. $W(n, \beta_k) \leq W(n, \beta_1)$. For $i \in \{0, \ldots, n - 1\}$ it is true that

$$x^{n-i} \leq 1 \iff (1-x^i)x^{n-i} \leq 1-x^i \iff x^{n-i} + x^i \leq 1 + x^n.$$  

Therefore

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

and

$$(1+x)\sum_{i=0}^{n-1} x^i = \frac{1}{2} \sum_{i=0}^{n-1} x^i + x^{i+1} + x^{n-i} + x^{n-i} \leq n(1+x^n).$$

Now

$$\frac{dW}{dx}(n, x) = -2 \frac{(1+x)\sum_{i=0}^{n-1} x^i - n(1+x^n)}{(1-x)^2} \geq 0$$

and the first inequality is shown. By

$$W(n, x) \leq \begin{cases} \frac{(n+1-x) - 2xn}{1-x} & x \in [-1, 0] \\ \frac{n+1-x}{1-x} & x \in (0, 1) \end{cases} \leq \frac{2n}{1-x},$$

the claim is proven.  

An explicit formula of the error if the initial state is chosen by the stationary distribution is established. Let us discuss the worst case error of $S_n$.

**Proposition 6.** Let $X_1, \ldots, X_n$ be a reversible Markov chain with respect to $\pi$, given by $(P, \pi)$. Then

$$\sup_{\|f\|_2 \leq 1} e(S_n, f)^2 = \frac{1 + \beta_1}{n(1 - \beta_1)} - \frac{2\beta_1(1 - \beta_1^n)}{n^2(1 - \beta_1)^2} \leq \frac{2}{n(1 - \beta_1)^2}.$$  

*Proof.* The individual error of $f$ is

$$e(S_n, f)^2 = \frac{1}{n^2} \sum_{k=1}^{|D|-1} |a_k|^2 W(n, \beta_k) \leq \frac{\|f\|_2^2}{n^2} \max_{k=1, \ldots, |D|-1} W(n, \beta_k)$$

$$= \frac{\|f\|_2^2}{n^2} W(n, \beta_1) = \frac{1 + \beta_1}{n(1 - \beta_1)} \|f\|_2^2 - \frac{2\beta_1(1 - \beta_1^n)}{n^2(1 - \beta_1)^2} \|f\|_2^2,$$
where \( a_k \) is chosen as in Proposition 4 and therefore \( \sum_{k=1}^{\left| D \right|} |a_k|^2 \leq \| f \|_2^2 \). From the preceding analysis of the individual error we have an upper bound. Now we consider \( f = u_1 \), where obviously \( \| u_1 \|_2 = 1 \) and get by applying (7) that
\[
e(S_n, u_1)^2 = \frac{1 + \beta_1}{n(1 - \beta_1)} - \frac{2\beta_1(1 - \beta_1^2)}{n^2(1 - \beta_1)^2}.
\]
Thus the error bound is attained for \( u_1 \) and by (8) everything is shown. \( \square \)

Finally an explicit presentation for the worst case error on the class of bounded functions with respect to \( \| \cdot \|_2 \) is shown. Notice, that (9) is an equality, which means that the integration error is completely known if we start with the stationary distribution. In some artificial cases this method even beats direct simulation, e.g. if all eigenvalues are smaller than zero or if one specific \( \beta_i < 0 \) and the goal is to approximate \( S(u_i) \). In [PHY92, Remark 3, p.617] the authors state a simple transition matrix where \( \beta_i = -\frac{1}{|D| - 1} \) for all \( i \). Now one could think to construct a transition matrix where \( \beta_1 \) is close to \(-1\) and therefore damp the integration error. But it is well known that this is not possible for large \( |D| \), since \( \beta_1 \geq -\frac{1}{|D| - 1} \).

In the next subsection we link the results to a more general framework, where the unrealistic assumption that the initial distribution is the stationary one is abandoned.

3.2. Starting from somewhere else. In the next statement a relation between the error of starting by \( \pi \) and the error of starting not by the invariant distribution is established.

**Proposition 7.** Let \( f \in \mathbb{R}^D \) and \( g := f - S(f) \). Let \( X_1, \ldots, X_{n+n_0} \) be a reversible Markov chain with respect to \( \pi \), given by \((P, \nu)\). Then
\[
e_\nu(S_{n,n_0}, f)^2 = e_\pi(S_n, f)^2 + \frac{1}{n^2} \sum_{j=1}^{n} L_{j+n_0}(g^2) + \frac{2}{n^2} \sum_{j=1}^{n} \sum_{k=j+1}^{n} L_{j+n_0}(gP^{k-j}g),
\]
where
\[
L_i(h) = \sum_{x \in D} d_i(x) h(x) \pi(x) = \sum_{x \in D} \sum_{y \in D} \frac{\nu(y)}{\pi(y)} (p^i(x, y) - \pi(y)) h(x) \pi(x).
\]

**Remark 3.** The proof of this identity is similar as in [Rud09], except for the fact that we study a discrete state space and therefore integrals become sums.
Proof. It is easy to see, that
\[
E_{\nu,P}|S(f) - S_{n,n_0}(f)|^2 = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} E_{\nu,P}(g(X_{n_0+j})g(X_{n_0+i}))
\]
\[
= \frac{1}{n^2} \sum_{j=1}^{n} \sum_{x \in D} P^{n_0+j}g^2(x) \nu(x) + \frac{2}{n^2} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \sum_{x \in D} P^{n_0+j}(gP^k-g)(x) \nu(x).
\]

For every function \( h \in \mathbb{R}^D \) and \( i \in \mathbb{N} \) under applying the reversibility the following transformation holds true
\[
\sum_{x \in D} (P^i h)(x) \nu(x) = \sum_{x \in D} \sum_{y \in D} h(y) p^i(x,y) \frac{\nu(x)}{\pi(x)} \pi(x)
\]
\[
= \text{rev.} \sum_{x \in D} \sum_{y \in D} \nu(y) \frac{\nu(y)}{\pi(y)} p^i(x,y) \pi(x)
\]
\[
= \sum_{x \in D} h(x) \pi(x) + \sum_{x \in D} \sum_{y \in D} \nu(y) \frac{\nu(y)}{\pi(y)} (p^i(x,y) - \pi(y)) h(x) \pi(x)
\]
\[
= \text{rev.} \sum_{x \in D} (P^i h)(x) \pi(x) + \sum_{x \in D} \sum_{y \in D} \nu(y) \frac{\nu(y)}{\pi(y)} (p^i(x,y) - \pi(y)) h(x) \pi(x).
\]

Using this in the setting above, formula (10) is shown. □

Equation (10) is still an error characterization where equality holds.

We will estimate \( L_k(h) \) to derive an upper bound. This depends very much on the speed of convergence from the chain to stationarity.

Lemma 8. Let \( h \in \mathbb{R}^D \), let again \( \beta = \max \{ \beta_1, |\beta|_{D|-1}| \} \). Then
\[
|L_k(h)| \leq \beta^k \sqrt{\frac{1}{\pi}} \frac{1}{\pi} \sqrt{\|\frac{\nu}{\pi} - 1\|_\infty} \cdot \|h\|_1,
\]
(11)
\[
|L_k(h)| \leq \beta^k \sqrt{\frac{\nu}{\pi} - 1} \|h\|_2.
\]
(12)

Proof. Let us consider \( L_k(h) = \langle d_k, h \rangle_\pi \). After applying Cauchy-Schwarz inequality we obtain
\[
|L_k(h)| \leq \|d_k\|_2 \|h\|_2.
\]
By applying (3) we showed (12). Inequality (12) and \( \|h\|_2 \leq \sqrt{\|\frac{1}{\pi}\|_\infty} \|h\|_1 \) implies (11). □

The ingredients for getting an explicit error bound for \( S_{n,n_0} \) are gathered together. Mainly the last Lemma ensures an exponential decay of \( L_k(h) \) which is used in the next Proposition.
Proposition 9. Let $X_1, \ldots, X_{n+n_0}$ be a reversible Markov chain with respect to $\pi$, given by $(P, \nu)$. Let $f \in \mathbb{R}^D$, $g := f - S(f)$ and

$$V(\beta, n) = \sum_{j=1}^{n} \beta^j + 2 \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \beta^k,$$

$$U(\beta, n) = \sum_{j=1}^{n} \left( \beta^j + 4 \sqrt{2} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \beta^k \right).$$

(i) Then for $g \in l^0_2$ we have

$$e_\nu(S_{n,n_0}, f)^2 \leq e_\pi(S_n, f)^2 + \frac{V(\beta, n)}{n^2} \beta^{n_0} \sqrt{\frac{1}{\pi} \left\| \frac{\nu}{\pi} - 1 \right\|_\infty} \|g\|_2^2.$$

(ii) Then for $g \in l^0_4$ we have

$$e_\nu(S_{n,n_0}, f)^2 \leq e_\pi(S_n, f)^2 + \frac{U(\beta, n)}{n^2} \beta^{n_0} \sqrt{\frac{1}{\pi} \left\| \frac{\nu}{\pi} - 1 \right\|_\infty} \|g\|_4^2.$$

(iii) Then for $g \in l^0_\infty$ we have

$$e_\nu(S_{n,n_0}, f)^2 \leq e_\pi(S_n, f)^2 + \frac{V(\beta, n)}{n^2} \beta^{n_0} \sqrt{\frac{1}{\pi} \left\| \frac{\nu}{\pi} - 1 \right\|_\infty} \|g\|_\infty^2.$$

Proof. As we have seen in (10) the error obeys

$$e_\nu(S_{n,n_0}, f)^2 = \frac{1}{n^2} \sum_{j=1}^{n} L_{j+n_0}(g^2) + \frac{2}{n^2} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} L_{j+n_0}(g_{P^k-g}).$$

Then by (11), Cauchy-Schwarz inequality and $\|P^{k-j}\|_{l^0_2-l^0_2} = \beta^{k-j}$ we get

$$|L_{j+n_0}(g^2)| \leq \sqrt{\frac{1}{\pi} \left\| \frac{\nu}{\pi} - 1 \right\|_\infty} \beta^{j+n_0} \|g\|_2^2,$$

$$|L_{j+n_0}(g_{P^k-g})| \leq \sqrt{\frac{1}{\pi} \left\| \frac{\nu}{\pi} - 1 \right\|_\infty} \beta^{k+n_0} \|g\|_2^2.$$
Putting this in the sums of equation (13) and let \( \varepsilon_0 = \sqrt{\frac{1}{n} \parallel g \parallel_\infty \sqrt{\parallel \frac{\nu}{\pi} - 1 \parallel_\infty \beta^{n_0}}} \) we obtain

\[
\sum_{j=1}^{n} |L_{j+n_0}(g^2)| + 2 \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} |L_{j+n_0}(g^{k-j}g)| \\
\leq \varepsilon_0 \parallel g \parallel_2^2 \sum_{j=1}^{n} \beta^j + \varepsilon_0 \parallel g \parallel_2^2 \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} 2 \beta^k \\
= \varepsilon_0 \parallel g \parallel_2^2 \left( \sum_{j=1}^{n} \beta^j + \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} 2 \beta^k \right) = V(\beta, n) \cdot \varepsilon_0 \parallel g \parallel_2^2.
\]

Thus claim (i) is shown. Now we use (12) and

\[
\|g^{k-j}g\|_2 \leq \|g\|_\infty \|g^{k-j}g\|_2 \leq \|g\|_\infty^2 \|P^{k-j}g\|_\infty \|P^{k-j}g\|_4 \leq \|g\|_\infty^2 \beta^{k-j}
\]

to obtain

\[
|L_{j+n_0}(g^2)| \leq \sqrt{\parallel \frac{\nu}{\pi} - 1 \parallel_\infty \beta^{j+n_0} \|g\|_\infty^2}, \\
|L_{j+n_0}(g^{k-j}g)| \leq \sqrt{\parallel \frac{\nu}{\pi} - 1 \parallel_\infty \beta^{k+n_0} \|g\|_\infty^2}.
\]

Exactly the same steps as in the proof of (i) follow, except for a different \( \varepsilon_0 = \sqrt{\parallel \frac{\nu}{\pi} - 1 \parallel_\infty \beta^{n_0}} \) and the supremum norm, i.e. assertion (iii) is proven. Let us turn to (ii). Again we use (12) and estimate

\[
\|g^{k-j}g\|_2 \leq \|g\|_4 \|g^{k-j}g\|_4 \leq \|g\|_4 \|g^{k-j}g\|_4 \leq 2\sqrt{2} \parallel g \parallel_4^2 \beta^{k-j}. \\
\]

Thus

\[
|L_{j+n_0}(g^2)| \leq \sqrt{\parallel \frac{\nu}{\pi} - 1 \parallel_\infty \beta^{j+n_0} \|g\|_4^2}, \\
|L_{j+n_0}(g^{k-j}g)| \leq 2\sqrt{2} \sqrt{\parallel \frac{\nu}{\pi} - 1 \parallel_\infty \beta^{k+n_0} \|g\|_4^2}.
\]
For $\varepsilon_0 = \sqrt{\frac{2}{\pi} - 1}\|\pi\|_{\infty}\beta_0$ we obtain

$$\sum_{j=1}^{n} |L_{j+n_0}(g^2)| + 4\sqrt{2} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} |L_{j+n_0}(gP^{k-j}g)|$$

$$\leq \varepsilon_0 \|g\|_{2}^2 \sum_{j=1}^{n} \beta^j + \varepsilon_0 \|g\|_{2}^2 \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} 4\sqrt{2} \beta^{\frac{k+j}{2}}$$

$$= \varepsilon_0 \|g\|_{2}^2 \left(\sum_{j=1}^{n} \beta^j + 4\sqrt{2} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \beta^{\frac{k+j}{2}}\right) = U(\beta, n) \cdot \varepsilon_0 \|g\|_{2}^2.$$

Finally by substituting this in equation (13) everything is shown. \(\square\)

In the last Proposition we introduced $V(\beta, n)$ and $U(\beta, n)$. These functions are bounded if $\beta < 1$. By applying the infinite geometric series several times the following is proven.

**Lemma 10.** For $n \in \mathbb{N}$ and $x \in [0, 1)$ we have

$$V(x, n) \leq \frac{2}{(1-x)^2}, \quad U(x, n) \leq \frac{4\sqrt{2}}{(1-x)(1-\sqrt{x})}.$$  

This implies that the asymptotic optimality is reached.

### 3.3. Main Theorem

**Theorem 11.** Let $X_1, \ldots, X_{n+n_0}$ be a reversible Markov chain with respect to $\pi$, given by $(P, \nu)$. Let $f \in \mathbb{R}^D$ and $a_k = \langle f, u_k \rangle_\pi$.

Then

$$\lim_{n \to \infty} n \cdot e_\nu(S_{n, n_0}, f)^2 = \lim_{n \to \infty} n \cdot e_\pi(S_n, f)^2 = \sum_{k=1}^{\lfloor D/2 \rfloor} |a_k|^2 \frac{1 + \beta_k}{1 - \beta_k}.$$  

(i) If we consider $f \in l_2$ then

$$e_\nu(S_{n, n_0}, f)^2 \leq \frac{2}{n(1 - \beta_1)} \|f\|_2^2 + \frac{2\sqrt{\|\nu\|_{\infty}\|\nu\|_{\infty} - 1}\|\pi\|_{\infty}\beta_0}{n^2(1 - \beta)^2} \|f\|_2^2.$$

(ii) If we consider $f \in l_4$ then

$$e_\nu(S_{n, n_0}, f)^2 \leq \frac{2}{n(1 - \beta_1)} \|f\|_4^2 + \frac{16\sqrt{2}\sqrt{\|\nu\|_{\infty} - 1}\|\pi\|_{\infty}\beta_0}{n^2(1 - \beta)(1 - \sqrt{\beta})} \|f\|_4^2.$$
(iii) If we consider $f \in l_\infty$ then
\[
e_\nu(S_{n,n_0}, f)^2 \leq \frac{2}{n(1 - \beta_1)} \|f\|_\infty^2 + \frac{4\sqrt{\|\nu - \pi\|_\infty \beta_0}}{n^2(1 - \beta)^2} \|f\|_\infty^2.
\]

Proof. By (10) and the fact that the remaining terms are going quadratic to zero as $n$ goes to infinity, we see that the asymptotic result holds true. For $f \in l_2$ we have $\|f - S(f)\|_2 \leq \|f\|_2$ and furthermore if $p \neq 2$ then
\[
\|f - S(f)\|_p \leq \|f\|_p + \|S(f)\|_p \leq \|f\|_p + \|f\|_1 \leq 2 \|f\|_p.
\]

Thus, via Proposition 9, Proposition 6 and Lemma 10 everything is shown. □

Notice, that from the estimate of Proposition 9 it follows immediately that
\[
\lim_{n \to \infty} n \cdot e_\nu(S_{n,n_0}, f)^2 \leq \lim_{n \to \infty} n \cdot e_\pi(S_{n,n_0}, f)^2 \leq \frac{1 + \beta_1}{1 - \beta_1} \|f\|_2^2.
\]

Thus there is no gap between the estimate and the asymptotical behavior. Also notice, that the upper bounds are continuous in the sense that if the initial distribution $\nu$ is $\pi$ then we obtain the bound of Proposition 6. The dependence of the bounds of (ii) and (iii) in Theorem 11 on the initial distribution is encouraging for an extension to general state spaces. (For an introduction to MCMC on general state spaces we refer to [RR04].) But the dependence of the initial distribution on the estimate in the $l_2$-case is disillusioning because of the additional factor of $\|\pi\|_\infty$.

In [Rud09, Theorem 8, p.10] a similar $l_\infty$-bound of $S_{n,n_0}$ for general state spaces is developed. This result holds for lazy, reversible Markov chains and may also be applied in the present setting, i.e. if the state space is discrete. In [Rud09] the asymptotic error limit is not attained. Thus we could improve the error bound and weaken the laziness condition, i.e. it is enough that $\beta_1 = \beta$. In [LPW09, Thm. 12.19, p.165] the authors obtained for another error term a comparable bound where the chain starts deterministically.

4. Burn-in

Let us assume that computer resources for the MCMC method for $N$ time steps are available, i.e. $N = n + n_0$. We want to choose the burn-in $n_0$ and the number of $n$ such that the error is as small as possible. The burn-in $n_0$ should be large but this implies that $n$ is possibly quite
small depending on how much resources we have. On the other hand $n$ should be large which again implies that $n_0$ is possibly small. There is obviously a trade-off between choosing the parameters. In the next statement we consider the error for an explicitly given burn-in, where for simplicity $\beta_1 = \beta$.

**Corollary 12.** Let $f \in \mathbb{R}^D$ be given and let

$$n_0 = \max \left\{ \left\lfloor \frac{\log(C)}{\log(\beta-1)} \right\rfloor, 0 \right\}.$$

(i) Let $C = \sqrt{\left\| \frac{1}{\pi} \right\|_\infty \sqrt{\left\| \frac{\nu}{\pi} - 1 \right\|_\infty}$, then

$$\sup_{\|f\|_2 \leq 1} e_\nu(S_{n,n_0}, f)^2 \leq \frac{2}{n(1-\beta)} + \frac{2}{n^2(1-\beta)^2}.$$ 

(ii) Let $C = 16\sqrt{2}\sqrt{\left\| \frac{\nu}{\pi} - 1 \right\|_\infty}$, then

$$\sup_{\|f\|_4 \leq 1} e_\nu(S_{n,n_0}, f)^2 \leq \frac{2}{n(1-\beta)} + \frac{1}{n^2(1-\beta)(1-\sqrt{\beta})}.$$ 

(iii) Let $C = 2\sqrt{\left\| \frac{\nu}{\pi} - 1 \right\|_\infty}$, then

$$\sup_{\|f\|_\infty \leq 1} e_\nu(S_{n,n_0}, f)^2 \leq \frac{2}{n(1-\beta)} + \frac{2}{n^2(1-\beta)^2}.$$ 

Note, that in the $l_\infty$- and $l_2$-case the error bound is the same. Just the constant $C$ which comes in by the density is different. This suggestion of the burn-in is justified in the following.

**4.1. Numerical experiments.** Suppose $C$ (very large), $\beta$ (close to one) and resources $N$ are given. The worst case error for $\|f\|_2 \leq 1$ or $\|f\|_\infty \leq 1$ is bounded by

$$b_\infty(n, n_0) := \sqrt{\frac{2}{n(1-\beta)} + \frac{2C\beta^{n_0}}{n(1-\beta)^2}}$$

and if we consider $\|f\|_4 \leq 1$ it is bounded by

$$b_4(n, n_0) := \sqrt{\frac{2}{n(1-\beta)} + \frac{C\beta^{n_0}}{n(1-\beta)(1-\sqrt{\beta})}}.$$ 

Since $N = n + n_0$ we can compute with a numerical procedure (here using Maple) the optimal choice of the burn-in denoted by $n_{\text{opt}}^4$, $n_{\text{opt}}^\infty$ to minimize the upper error bounds. (This is a simple one dimensional minimization problem with different parameters.)
ERROR BOUNDS OF MCMC

Table 1. For $C = 10^{30}$ where $n_{opt}^i$ minimizes

$$b_i(N - n_{opt}^i, n_{opt}^i), \ i = 4, \infty.$$ 

Table 1 gives a collection of typical results. It turned out that the
above suggested lower bound is almost the optimal choice. The com-
puted value $n_{opt}^4$ and $n_{opt}^\infty$ is almost the same as $n_0 = \lceil \log(C) / \log(\beta^{-1}) \rceil$.
In the case $C = 10^{30}$ and $\beta = 0.999$ Theorem 11 gives for no choice of $n$ and $n_0$ with $N = 10^4$ an error smaller than one.

In Figure 1 we plotted $b_i(N - n_0, n_0)$ for different $n_0$ and $e_\pi(S_N, u_1) = \sqrt{\frac{1+\beta_1}{N(1-\beta_1)} - \frac{2\beta_1(1-\beta^N)}{N^2(1-\beta_1)^2}}$. Roughly spoken one may see in Figure 1 that if the burn-in is chosen too small a vertical shifting takes place and if the burn-in is chosen to large a horizontal shifting takes place. The

![Figure 1](image.png)
asymptotic behavior is the same, i.e. for the long run the error of \( S_{n,n_0} \) converges to the error of \( S_n \). If \( \beta \) and \( C \) are given we chose the burn-in as suggested above. If there is an estimate of \( \log(C)/\log(\beta^{-1}) \) one should ensure that it is not smaller than the real ratio. As seen in Figure 1 if it is slightly smaller there is already strong influence. By choosing the burn-in too large the influence is less heavy.

Finally if there is no estimation or computation of the parameters \( \beta \) or \( C \) a simple but very efficient strategy is given by choosing \( n = n_0 = \frac{N}{2} \) (for even \( N \)). In Figure 2 we see \( b_4(N, \frac{N}{2}), b_4(N - n_0, n_0) \) and \( e_\pi(S_N, u_1) \). In the asymptotic behavior we pay the price of a factor of \( \sqrt{2} \), i.e. the asymptotic error is \( \sqrt{2} \) times larger than \( e_\pi(S_N, u_1) \) where we started in equilibrium. This strategy works well and reaches the same convergence rate as choosing the burn-in as suggested above, which is seen in Figure 2.

**Acknowledgements**

The author thanks Erich Novak and Aicke Hinrichs for their valuable comments.
ERROR BOUNDS OF MCMC

REFERENCES


FRIEDRICH SCHILLER UNIVERSITY JENA, MATHEM. INSTITUTE, ERNST-ABBE-PLATZ 2, D-07743 JENA, GERMANY

E-mail address: daniel.rudolf@uni-jena.de
Preprint Series DFG-SPP 1324

http://www.dfg-spp1324.de

Reports


