Computation of expectations by Markov chain Monte Carlo methods

Erich Novak and Daniel Rudolf

Abstract Markov chain Monte Carlo (MCMC) methods are a very versatile and widely used tool to compute integrals and expectations. In this short survey we focus on error bounds, rules for choosing the burn in, high dimensional problems and tractability versus curse of dimension.

1 Motivation

Consider the following example. We want to compute

$$\mathbb{E}_G(f) = \frac{1}{\text{vol}_d(G)} \int_G f(x) \, dx,$$

where $f$ belongs to some class of functions and $G$ belongs to some class of sets. We assume that $G \subset \mathbb{R}^d$ is measurable with $0 < \text{vol}_d(G) < \infty$, where $\text{vol}_d$ denotes the Lebesgue measure. Thus, we want to compute the expected value of $f$ with respect to the uniform distribution on $G$.

The input $(f, G)$ is given by an oracle: For $x \in G$ we can compute $f(x)$ and $G$ is given by a membership oracle, i.e. we are able to check whether any $x \in \mathbb{R}^d$ is in $G$ or not. We always assume that $G$ is convex and will work with the class

$$\mathcal{G}_{r,d} = \{ G \subset \mathbb{R}^d : G \text{ is convex, } B_d \subset G \subset rB_d \},$$

where $r \geq 1$ and $rB_d = \{ x \in \mathbb{R}^d : |x| \leq r \}$ is the Euclidean ball with radius $r$.  

Erich Novak
Friedrich Schiller University Jena, Mathematical Institute, Ernst-Abbe-Platz 2, D-07743 Jena, Germany, e-mail: erich.novak@uni-jena.de

Daniel Rudolf
Friedrich Schiller University Jena, Mathematical Institute, Ernst-Abbe-Platz 2, D-07743 Jena, Germany, e-mail: daniel.rudolf@uni-jena.de
A first approach might be a simple acceptance/rejection method. The idea is to generate a point in \( rB_d \) according to the uniform distribution and if it is in \( G \) it is accepted, otherwise it is rejected. If \( x_1, \ldots, x_n \in G \) are the accepted points then we output the mean value of the \( f(x_i) \). However, this method does not work reasonably since the acceptance probability can be extremely small, it can be \( r^{-d} \).

It seems that all known efficient algorithms for this problem use Markov chains. The idea is to find a sampling procedure that approximates a sample with respect to the uniform distribution in \( G \). More precisely, we run a Markov chain to approximate the uniform distribution for any \( G \in \mathcal{G}_{r,d} \). Let \( X_1, X_2, \ldots, X_{n+n_0} \) be the first \( n+n_0 \) steps of such a Markov chain. Then

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

is an approximation of \( \mathbb{E}_G(f) \). The additional parameter \( n_0 \) is called burn-in and, roughly spoken, is the number of steps of the Markov chain to get close to the uniform distribution.

## 2 Approximation of expectations by MCMC

### 2.1 Preliminaries

We provide the basics of Markov chains. For further reading we refer to the paper [14] of Roberts and Rosenthal which surveys various results about Markov chains on general state spaces.

A Markov chain is a sequence of random variables \((X_n)_{n \in \mathbb{N}}\) which satisfies the Markov property. For \( i \in \mathbb{N} \), the conditional distribution of \( X_{i+1} \) depends only on \( X_i \) and not on \((X_1, \ldots, X_{i-1})\),

\[
P(X_{i+1} \in A \mid X_1, \ldots, X_i) = P(X_{i+1} \in A \mid X_i).
\]

By \( \mathcal{B}(G) \) we denote the Borel \( \sigma \)-algebra of \( G \). Let \( \nu \) be a distribution on \((G, \mathcal{B}(G))\) and let \( K: G \times \mathcal{B}(G) \to [0,1] \) be a transition kernel, i.e. \( K(x, \cdot) \) is a probability measure for each \( x \in G \) and \( K(\cdot, A) \) is a \( \mathcal{B}(G) \)-measurable real-valued function for each \( A \in \mathcal{B}(G) \). A transition kernel and a distribution \( \nu \) give rise to a Markov chain \((X_n)_{n \in \mathbb{N}}\) in the following way. Assume that the distribution of \( X_1 \) is given by \( \nu \). Then, for \( i \geq 2 \) and a given \( X_{i-1} = x_{i-1} \), we have \( X_i \) with distribution \( K(x_{i-1}, \cdot) \), that is, for all \( A \in \mathcal{B}(G) \), the conditional probability that \( X_i \in A \) is given by \( K(x_{i-1}, A) \). We call such a sequence of random variables a Markov chain with transition kernel \( K \) and initial distribution \( \nu \).

In the whole paper we only consider Markov chains with reversible transition kernel, we assume that there exists a probability measure \( \pi \) on \( \mathcal{B}(G) \) such that
In particular any such \( \pi \) is a stationary distribution of \( K \), i.e.,

\[
\pi(A) = \int_G K(x, A) \pi(dx), \quad A \in \mathcal{B}(G).
\]

Further, the transition kernel induces an operator on functions and an operator on measures given by

\[
P f(x) = \int_G f(y) K(x, dy), \quad \text{and} \quad \nu P(A) = \int_G K(x, A) \nu(dx),
\]

where \( f \) is \( \pi \)-integrable and \( \nu \) is absolutely continuous with respect to \( \pi \). One has

\[
\mathbb{E}[f(X_n) | X_1 = x] = P^{n-1} f(x) \quad \text{and} \quad \mathbb{P}\nu(X_n \in A) = \nu P^{n-1}(A),
\]

for \( x \in G, A \in \mathcal{B}(G) \) and \( n \in \mathbb{N} \), where \( \nu \) in \( \mathbb{P}\nu \) indicates that \( X_1 \) has distribution \( \nu \). By the reversibility with respect to \( \pi \) we have \( \frac{d\nu^P}{d\pi}(x) = P(\frac{d\nu}{d\pi})(x) \), where \( \frac{d\nu}{d\pi} \) indicates that \( \nu \) has distribution \( \pi \) with respect to \( \pi \).

Further, for \( p \in [1, \infty) \) let \( L_p = L_p(\pi) \) be the space of measurable functions \( f : G \to \mathbb{R} \) which satisfy

\[
||f||_p = \left( \int_G |f(x)|^p \pi(dx) \right)^{1/p} < \infty.
\]

The operator \( P : L_p \to L_p \) is linear and bounded and by the reversibility \( P : L_2 \to L_2 \) is self-adjoint.

The goal is to quantify the speed of convergence, if it converges at all, of \( \nu P^n \) to \( \pi \) for increasing \( n \in \mathbb{N} \). For this we use the total variation distance between two probability measures \( \nu, \mu \) on \( (G, \mathcal{B}(G)) \) given by

\[
\|\nu - \mu\|_{tv} = \sup_{A \in \mathcal{B}(G)} |\nu(A) - \mu(A)|.
\]

It is helpful to consider the total variation distance as an \( L_1 \)-norm, see for example [14].

\textbf{Lemma 1.} Assume the probability measures \( \nu, \mu \) have densities \( \frac{d\nu}{d\pi}, \frac{d\mu}{d\pi} \in L_1 \), then

\[
\|\nu - \mu\|_{tv} = \frac{1}{2} \left\| \frac{d\nu}{d\pi} - \frac{d\mu}{d\pi} \right\|_1.
\]

Now we ask for an upper bound of \( \|\nu P^n - \pi\|_{tv} \).

\textbf{Lemma 2.} Let \( \nu \) be a probability measure on \( (G, \mathcal{B}(G)) \) with \( \frac{d\nu}{d\pi} \in L_1 \) and let \( S(f) = \int_G f(x) \pi(dx) \). Then, for any \( n \in \mathbb{N} \) holds

\[
\|\nu P^n - \pi\|_{tv} \leq \|P^n - S\|_{L_1 \to L_1} \leq \left( \frac{1}{2} \left\| \frac{d\nu}{d\pi} - 1 \right\|_1 \right) \leq \|P^n - S\|_{L_1 \to L_1}
\]
and
\[ \|vP^n - \pi\|_{tv} \leq \|P^n - S\|_{L^2 \rightarrow L^2}\frac{1}{2} \left\| \frac{d\nu}{d\pi} - 1 \right\|_2. \]

Proof. By Lemma 1, by \( P^n 1 = 1 \) and by the reversibility, in particular \( \frac{d(\nu P^n)}{d\pi}(x) = P^n(\frac{d\nu}{d\pi})(x) \), we have
\[ 2 \|vP^n - \pi\|_{tv} = \left\| \frac{d(\nu P^n)}{d\pi} - 1 \right\|_1 = \left\| P^n \left( \frac{d\nu}{d\pi} - 1 \right) \right\|_1 = \left\| (P^n - S) \left( \frac{d\nu}{d\pi} - 1 \right) \right\|_1. \]

Note that the last equality comes from \( S(\frac{d\nu}{d\pi} - 1) = 0 \).

Observe that for \( \nu = \pi \) the left-hand side and also the right-hand side of the estimates are zero.

Let us consider \( \|P^n - S\|_{L^2 \rightarrow L^2} \). Because of the reversibility with respect to \( \pi \) we obtain the following, see for example [19, Lemma 3.16, p. 45].

**Lemma 3.** For \( n \in \mathbb{N} \) we have
\[ \|P^n - S\|_{L^2 \rightarrow L^2} = \|(P - S)^n\|_{L^2 \rightarrow L^2} = \|P - S\|_{L^2 \rightarrow L^2}^n. \]

The last two lemmata motivate the following two convergence properties of transition kernels.

**Definition 1 (L1-exponential convergence).** Let \( \alpha \in [0, 1) \) and \( M \in (0, \infty) \). Then the transition kernel \( K \) is \( L^1 \)-exponentially convergent with \((\alpha, M)\) if
\[ \|P^n - S\|_{L^1 \rightarrow L^1} \leq \alpha^n M, \quad n \in \mathbb{N}. \quad (2) \]

A Markov chain with transition kernel \( K \) is called \( L^1 \)-exponentially convergent if there exist an \( \alpha \in [0, 1) \) and \( M \in (0, \infty) \) such that (2) holds.

**Definition 2 (L2-spectral gap).** We say that a transition kernel \( K \) and its corresponding Markov operator \( P \) have an \( L^2 \)-spectral gap if
\[ \text{gap}(P) = 1 - \|P - S\|_{L^2 \rightarrow L^2} > 0. \]

If the transition kernel has an \( L^2 \)-spectral gap, then by Lemma 2 and Lemma 3 we have that
\[ \|vP^n - \pi\|_{tv} \leq (1 - \text{gap}(P))^n \left\| \frac{d\nu}{d\pi} - 1 \right\|_2. \]

Next, we define other convergence properties which are based on the total variation distance.

**Definition 3 (uniform ergodicity and geometric ergodicity).** Let \( \alpha \in [0, 1) \) and \( M : G \rightarrow (0, \infty) \). Then the transition kernel \( K \) is called geometrically ergodic with \((\alpha, M(x))\) if one has for \( \pi \)-almost all \( x \in G \) that
Computation of expectations by Markov chain Monte Carlo methods

\[ \|K^n(x, \cdot) - \pi\|_{tv} \leq M(x) \alpha^n, \quad n \in \mathbb{N}. \]  
(3)

If the inequality (3) holds with a bounded function \( M(x) \), i.e.

\[ \sup_{x \in G} M(x) \leq M' < \infty, \]

then \( K \) is called uniformly ergodic with \((\alpha, M')\).

Now we state several relations between the different properties. Since we assume that the transition kernel is reversible with respect to \( \pi \) we have the following:

\[
\begin{align*}
\text{uniformly ergodic} & \iff L_1\text{-exponentially convergent} \\
& \text{with } (\alpha, M) \quad \text{with } (\alpha, 2M) \\
& \Downarrow \\
\text{geometrically ergodic} & \iff L_2\text{-spectral gap } \geq 1 - \alpha \\
& \text{with } (\alpha, M(x)) \quad \text{with } (\alpha, M(x)) \quad 1 - \alpha.
\end{align*}
\]
(4)

The fact that uniform ergodicity implies geometric ergodicity is obvious. For the proofs of the other relations and further details we refer to [19, Proposition 3.23, Proposition 3.24]. Further, if the transition kernel is \( \phi \)-irreducible, for details we refer to [13] and [15], then

\[
\text{geometrically ergodic } \iff L_2\text{-spectral gap } \geq 1 - \alpha. \]
(5)

2.2 Mean square error bounds of MCMC

The goal is to compute

\[ S(f) = \int_G f(x) \pi(dx). \]

We use an average of a finite Markov chain sample as approximation of the mean, i.e. we approximate \( S(f) \) by

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

The number \( n \) determines the number of function evaluations of \( f \). The number \( n_0 \) is the burn-in or warm up time. Intuitively, it is the number of steps of the Markov chain to get close to the stationary distribution \( \pi \).

We study the mean square error of \( S_{n,n_0} \), given by

\[ e_V(S_{n,n_0}, f) = (E_{V,K}[S_{n,n_0}(f) - S(f)])^{1/2}, \]
where $\nu$ and $K$ indicate the initial distribution and transition kernel. We start with the case $\nu = \pi$, where the initial distribution is the stationary distribution.

**Lemma 4.** Let $(X_n)_{n \in \mathbb{N}}$ be a Markov chain with transition kernel $K$ and initial distribution $\pi$. We define

$$\Lambda = \sup \{ \alpha : \alpha \in \text{spec}(P-S) \},$$

where $\text{spec}(P-S)$ denotes the spectrum of the operator $P-S : L_2 \to L_2$, and assume that $\Lambda < 1$. Then

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

For a proof of this result we refer to [19, Corollary 3.27]. Let us discuss the assumptions and implications of Lemma 4. First, note that for the simple Monte Carlo method we have $\Lambda = 0$. In this case we get (up to a constant of 2) what we would expect. Further, note that gap $(P) = 1 - \|P-S\|_{L_2 \to L_2}$ and

$$\|P-S\|_{L_2 \to L_2} = \sup \{ |\alpha| : \alpha \in \text{spec}(P-S) \},$$

so that gap $(P) \leq 1 - \Lambda$. This also implies that if $P : L_2 \to L_2$ is positive semidefinite we obtain gap $(P) = 1 - \Lambda$. Thus, whenever we have a lower bound for the spectral gap we can apply Lemma 4 and can replace $1 - \Lambda$ by gap $(P)$. Further note if $\gamma \in [0,1)$, $M \in (0, \infty)$ and the transition kernel is $L_1$-exponentially convergent with $(\gamma,M)$ then we have, using (4), that gap $(P) \geq 1 - \gamma$.

Now we ask how $e_\nu(S_{n,n_0},f)$ behaves depending on the initial distribution. The idea is to decompose the error in a suitable way. For example in a bias and variance term. However, we want to have an estimate with respect to $\|f\|_2$ and in this setting the following decomposition is more convenient:

$$e_\nu(S_{n,n_0},f)^2 = e_\pi(S_{n,n_0},f)^2 + \text{rest},$$

where rest denotes an additional term such that equality holds. Then, we estimate the remainder term and use Lemma 4 to obtain an error bound. For further details of the proof of the following error bound we refer to [19, Theorem 3.34 and Theorem 3.41].

**Theorem 1.** Let $(X_n)_{n \in \mathbb{N}}$ be a Markov chain with reversible transition kernel $K$ and initial distribution $\nu$. Further, let

$$\Lambda = \sup \{ \alpha : \alpha \in \text{spec}(P-S) \},$$

where $\text{spec}(P-S)$ denotes the spectrum of the operator $P-S : L_2 \to L_2$, and assume that $\Lambda < 1$. Then

$$\sup_{\|f\|_p \leq 1} e_\nu(S_{n,n_0},f)^2 \leq \frac{2}{n(1-\Lambda)} + \frac{2C_\nu \gamma^{11}}{n^2(1-\gamma)^2},$$

(6)

holds for $p = 2$ and for $p = 4$ under the following conditions
1. for \( p = 2 \), \( \frac{d\nu}{d\pi} \in L_\infty \) and a transition kernel \( K \) which is \( L_1 \)-exponentially convergent with \( (\gamma, M) \) where \( C_\nu = M \frac{\|\frac{d\nu}{d\pi} - 1\|_\infty}{\|\frac{d\nu}{d\pi} - 1\|_1} \);

2. for \( p = 4 \), \( \frac{d\nu}{d\pi} \in L_2 \) and \( 1 - \gamma = \text{gap}(P) > 0 \) where \( C_\nu = 64 \frac{\|\frac{d\nu}{d\pi} - 1\|_2}{\|\frac{d\nu}{d\pi} - 1\|_2} \).

Let us discuss the results. If the transition kernel is \( L_1 \)-exponentially ergodic, then we have an explicit error bound for integrands \( f \in L_2 \) whenever the initial distribution has a density \( \frac{d\nu}{d\pi} \in L_\infty \). However, in general it is difficult to provide explicit values \( \gamma \) and \( M \) such that the transition kernel is \( L_1 \)-exponentially convergent with \( (\gamma, M) \). This motivates to consider transition kernel which satisfy a weaker convergence property, such as the existence of an \( L_2 \)-spectral gap. In this case we have an explicit error bound for integrands \( f \in L_4 \) whenever the initial distribution has a density \( \frac{d\nu}{d\pi} \in L_2 \). Thus, by assuming a weaker convergence property of the transition kernel we obtain a weaker result in the sense that \( f \) must be in \( L_4 \) rather than \( L_2 \).

However, with respect to \( \frac{d\nu}{d\pi} \) we do not need boundedness anymore, it is enough that \( \frac{d\nu}{d\pi} \in L_2 \).

In Theorem 1 we provided explicit error bounds and we add in passing that also other error bounds are known, see [1, 4, 5, 19].

If we want to have an error of \( \varepsilon \in (0, 1) \) it is still not clear how to choose \( n \) and \( n_0 \) to minimize the total amount of steps \( n + n_0 \). How should we choose the burn-in \( n_0 \)? Let \( e(n, n_0) \) be the right hand side of (6) and assume that \( \Lambda = \gamma \). Further, assume that we have computational resources for \( N = n + n_0 \) steps of the Markov chain. We want to get an \( n_{\text{opt}} \) which minimizes \( e(N - n_0, n_0) \).

Let \( n_{\text{opt}} \) be the right hand side of (6) and assume that \( \Lambda = \gamma \). Further, assume that we have computational resources for \( N = n + n_0 \) steps of the Markov chain. We want to get an \( n_{\text{opt}} \) which minimizes \( e(N - n_0, n_0) \). In [19, Lemma 2.26] the following is proven: For all \( \delta > 0 \) and large enough \( N \) and \( \Lambda \), the number \( n_{\text{opt}} \) satisfies

\[
\frac{\log C_\nu}{\log \gamma - 1} \leq n_{\text{opt}} \leq \frac{\log C_\nu}{\log \gamma^{-1}},
\]

Further note that \( \log \gamma^{-1} \geq 1 - \gamma \). Thus, in this setting \( n_{\text{opt}} = \left\lfloor \frac{\log C_\nu}{1 - \gamma} \right\rfloor \) is a reasonable and almost optimal choice for the burn-in.

### 3 Application of the error bound and limitations of MCMC

First, we briefly introduce a technique to prove a lower bound of the spectral gap if the Markov operator of a transition kernel is positive semidefinite on \( L_2 \). The following result, known as Cheeger’s inequality, is in this form due to Lawler and Sokal [6].

**Proposition 1.** Let \( K \) be a reversible transition kernel, which induces a Markov operator \( P : L_2 \rightarrow L_2 \). Then

\[
\frac{\phi^2}{2} \leq 1 - \Lambda \leq 2\phi,
\]

where \( \Lambda = \sup\{\alpha : \alpha \in \text{spec}(P - S)\} \) and
\[ \phi = \inf_{0 < \pi(A) \leq 1/2} \frac{\int_A K(x, A^c) \pi(dx)}{\pi(A)} \]

is the conductance of \( K \).

Now we state different applications of Theorem 1.

### 3.1 Hit-and-run algorithm

We consider the example of Section 1. Let \( G \in \mathcal{G}_r, \) see (1), and let \( \mu_G \) be the uniform distribution in \( G \). We define

\[ \mathcal{F}_{r,d} = \{ (f, G) : G \in \mathcal{G}_r, f \in L_4(\mu_G), \|f\|_4 \leq 1 \} \]  

(7)

The goal is to approximate

\[ S(f, 1_G) = \frac{1}{\text{vol}(G)} \int_G f(x) \, dx, \]

where \( (f, G) \in \mathcal{F}_{r,d} \). The hit-and-run algorithm defines a Markov chain which satisfies the assumptions of Theorem 1. A step from \( x \in G \) of the hit-and-run algorithm works as follows

1. Choose a direction, say \( \theta \), uniformly distributed on the sphere \( \partial B_d \).
2. Choose the next state, say \( y \in G \), uniformly distributed in \( G \cap \{ x + \theta r : r \in \mathbb{R} \} \).

After choosing a direction \( \theta \) one samples the next state \( y \in G \) with respect to the uniform distribution in the line determined by the current state \( x \) and the direction \( \theta \) restricted to \( G \). The random number, say \( u \in [0, 1] \), for the second part is chosen independently of the first part and also all steps are independent.

Lovász and Vempala prove in [7, Theorem 4.2, p. 993] a lower bound of the conductance \( \phi \), see Proposition 1 for the definition of the conductance.

**Proposition 2.** Let \( G \in \mathcal{G}_r, \) Then, the conductance of the hit-and-run algorithm is bounded from below by \( 2^{-25}(dr)^{-1} \).

It is known that the hit-and-run algorithm induces a positive semidefinite Markov operator, say \( H \), see [17]. By Proposition 1 we obtain

\[ \text{gap}(H) \geq 2^{-51} \]

and Theorem 1 implies the following error bound for the class \( \mathcal{F}_{r,d} \), see (1) and (7).

**Theorem 2.** Let \( \nu \) be the uniform distribution on \( B_d \). Let \( (X_n)_{n \in \mathbb{N}} \) be a Markov chain with transition kernel, given by the hit-and-run algorithm, and initial distribution \( \nu \). Let
\[ n_0 = \left\lceil 4.51 \cdot 10^{15} d^2 r^2 (d \log r + 4.16) \right\rceil. \]

Then
\[
\sup_{(f,G) \in \mathcal{F}_{r,d}} e_V(S_{n,n_0}, (f,1_G)) \leq 9.5 \cdot 10^{-7} \frac{d r}{\sqrt{n}} + 6.4 \cdot 10^{15} d^2 r^2 n.
\]

This result states that the number of oracle calls for \( f \) and \( G \) to obtain an error \( \varepsilon > 0 \) is bounded by \( \kappa d^2 r^2 (\varepsilon^{-2} + d \log r) \), for an explicit constant \( \kappa > 0 \). Hence the computation of \( S(f,1_G) \) on the class \( \mathcal{F}_{r,d} \) is polynomially tractable, see [11, 11, 12]. The tractability result can be extended also to other classes of functions, see [18]. Note that we applied the second statement of Theorem 1. It is known that the hit-and-run algorithm is \( L_1 \)-exponentially ergodic with \( (\gamma, M) \), for some \( \gamma \in (0,1) \) and \( M \in (0,\infty) \). But the best known numbers \( \gamma \) and \( M \) are exponentially bad in terms of the dimension, see [20].

### 3.2 Metropolis-Hastings algorithm

Let \( G \subset \mathbb{R}^d \) and \( \rho : G \to (0,\infty) \), where \( \rho \) is integrable with respect to the Lebesgue measure. We define the distribution \( \pi_\rho \) on \( (G, \mathcal{B}(G)) \) by
\[
\pi_\rho(A) = \frac{\int_A \rho(x) \, dx}{\int_G \rho(x) \, dx}, \quad A \in \mathcal{B}(G).
\]

The goal is to compute
\[
S(f,\rho) = \int_G f(x) \pi_\rho(\,dx) = \frac{\int_G f(x) \rho(x) \, dx}{\int_G \rho(x) \, dx}
\]
for functions \( f : G \to \mathbb{R} \) which are integrable with respect to \( \pi_\rho \).

The **Metropolis-Hastings algorithm** defines a Markov chain which approximates \( \pi_\rho \). We need some further notations. Let \( q : G \times G \to [0,\infty] \) be a function such that \( q(x,\cdot) \) is Lebesgue integrable for all \( x \in G \) with \( \int_G q(x,y) \, dy \leq 1 \). Then
\[
Q(x,A) = \int_A q(x,y) \, dy + 1_A(x) \left( 1 - \int_G q(x,y) \, dy \right), \quad x \in G, A \in \mathcal{B}(G),
\]
is a transition kernel and we call \( q(\cdot,\cdot) \) \textit{transition density}. The idea is to modify \( Q \), such that \( \pi_\rho \) gets a stationary distribution of the modification. We propose a state with \( Q \) and with a certain probability, which depends on \( \rho \), the state is accepted. Let \( \alpha(x,y) \) be the acceptance probability
\[
\alpha(x,y) = \begin{cases} 
1 & \text{if } q(x,y) \rho(y) = 0, \\
\min\{1, \frac{q(y,x) \rho(y)}{q(x,y) \rho(x)}\} & \text{otherwise}.
\end{cases}
\]

The transition kernel of the Metropolis-Hastings algorithm is
\[ K_\rho(x,A) = \int_A \alpha(x,y) q(x,y) \, dy + I_A(x) \left( 1 - \int_G \alpha(x,y) q(x,y) \, dy \right) \]

for \( x \in G \) and \( A \in \mathcal{B}(G) \). The transition kernel \( K_\rho \) is reversible with respect to \( \pi_\rho \). From the current state \( x \in G \) a single transition of the algorithm works as follows:

1. Sample a proposal state \( y \in G \) with respect to \( Q(x,\cdot) \).
2. With probability \( \alpha(x,y) \) return \( y \), otherwise reject \( y \) and return \( x \).

Again, all steps are done independently of each other. If \( q(x,y) = q(y,x) \), i.e. \( q \) is symmetric, then \( K_\rho \) is called Metropolis algorithm and if \( q(x,y) = \eta(y) \) for a function \( \eta: G \to (0,\infty) \) for all \( x,y \in G \), then \( K_\rho \) is called independent Metropolis algorithm.

Let \( G \subset \mathbb{R}^d \) be bounded and for \( C \geq 1 \) let

\[ \mathcal{R}_C = \{ \rho: G \to (0,\infty) \mid 1 \leq \rho(x) \leq C \}. \quad (8) \]

Thus, for any \( \rho \in \mathcal{R}_C \) holds \( \sup \rho / \inf \rho \leq C \). If \( \rho: G \to (0,\infty) \) satisfies \( \sup \rho / \inf \rho \leq C \), then

\[ \frac{\| \rho \|_\infty}{C} \leq \rho(x) \leq C \inf \rho. \]

Thus, \( C \cdot \rho / \| \rho \|_\infty \in \mathcal{R}_C \). We consider an independent Metropolis algorithm. The proposal transition kernel is

\[ Q(x,A) = \mu_G(A) = \frac{\text{vol}_d(A)}{\text{vol}_d(G)}, \quad A \in \mathcal{B}(G), \]

i.e. a state is proposed with the uniform distribution in \( G \). Then

\[ K_\rho(x,A) = \int_A \alpha(x,y) \frac{dy}{\text{vol}_d(G)} + I_A(x) \left( 1 - \int_G \alpha(x,y) \frac{dy}{\text{vol}_d(G)} \right), \]

where \( \alpha(x,y) = \min \{ 1, \rho(y)/\rho(x) \} \). The transition operator \( P_\rho: L_2(\pi_\rho) \to L_2(\pi_\rho) \), induced by \( K_\rho \), is positive semidefinite. For details we refer to [17]. Thus, \( \text{gap}(P_\rho) = 1 - A_\rho \), with \( A_\rho = \Lambda \). Further, for \( \rho \in \mathcal{R}_C \) Theorem 2.1 of [9] provides a criterion for uniform ergodicity of the independent Metropolis algorithm. Namely, \( K_\rho \) is uniformly ergodic with \( (\gamma,1) \) for \( \gamma = 1 - C^{-1}/\text{vol}_d(G) \). Thus, by (4) we have that it is \( L_1 \)-exponentially ergodic with \( (\gamma,2) \). Further, by (4) we obtain

\[ 1 - A_\rho = \text{gap}(P_\rho) \geq \frac{C^{-1}}{\text{vol}_d(G)}. \]

Let

\[ \mathcal{F}_{C,d} = \{ (f,\rho) : \rho \in \mathcal{R}_C, f \in L_2(\pi_\rho), \| f \|_2 \leq 1 \}. \quad (9) \]

We apply Theorem [1] and obtain for the class \( \mathcal{F}_{C,d} \) (see (8) and (9))
Theorem 3. Let \((X_n)_{n \in \mathbb{N}}\) be a Markov chain with transition kernel, given by the Metropolis algorithm with proposal \(\mu_G\), and initial distribution \(\mu_G\). Let

\[ n_0 = \lceil C \text{vol}_d(G) \log(2C) \rceil. \]

Then

\[ \sup_{(f, \rho) \in \mathcal{F}_{C,d}} e_{\nu}(S_{n,n_0}, (f, \rho))^2 \leq \frac{2C \text{vol}_d(G) n}{n} + \frac{4C^2 \text{vol}_d(G)^2}{n^2}. \]

The upper bound in Theorem 3 does not depend on the dimension \(d\), as long as \(\text{vol}_d(G)\) and \(C\) do not depend on \(d\). In some applications, however, the upper bound is rather useless since \(C = C_d\) is exponentially large in \(d\). Assume, for example, that

\[ \rho(x) = \exp(-\alpha|x|^2), \quad (10) \]

i.e. \(\rho\) is the non-normalized density of a \(N(0, \sqrt{2\alpha}^{-1})\) random variable. We consider scaled versions of \(\rho\). If \(G = B_d\), then \(\exp(\alpha)\rho \in \mathcal{R}_{\exp(\alpha)}\) and if \(G = [-1,1]^d\), then \(\exp(\alpha d)\rho \in \mathcal{R}_{\exp(\alpha d)}\). This is bad, since \(C\), for example \(\exp(\alpha)\) or \(\exp(\alpha d)\), might depend exponentially on \(\alpha\) and \(d\).

This example shows that we would greatly prefer an upper bound where \(C\) is replaced by a power of \(\log C\). However, on the class \(\mathcal{F}_{C,d}\) this is not possible. The same proof as in [3] Theorem 1] leads to the following lower bound for all randomized algorithms.

Theorem 4. Any randomized algorithm \(S_n\) that uses \(n\) values of \(f\) and \(\rho\) satisfies the lower bound

\[ \sup_{(f, \rho) \in \mathcal{F}_{C,d}} e(S_n, (f, \rho)) \geq \frac{\sqrt{2}}{6} \left\{ \begin{array}{ll} \sqrt{\frac{C}{2n}} & \quad 2n \geq C - 1, \\ \frac{3C}{C + 2n - 1} & \quad 2n < C - 1. \end{array} \right. \]

The class \(\mathcal{F}_{C,d}\) is too large. Thus the error bound is not satisfying. In the following we prove a much better upper bound for a smaller class of densities. Let \(G = B_d\) and let \(\rho\) be log-concave, i.e. for all \(\lambda \in (0, 1)\) and for all \(x, y \in B_d\) we have

\[ \rho(\lambda x + (1-\lambda)y) \geq \rho(x)^\lambda \rho(y)^{1-\lambda}. \quad (11) \]

Then let

\[ \mathcal{R}_{\alpha,d} = \{ \rho : B_d \to (0, \infty) \mid \rho \text{ is log-concave, } |\log \rho(x) - \log \rho(y)| \leq \alpha |x-y| \}. \quad (12) \]

We consider log-concave densities where \(\log \rho\) is Lipschitz continuous with constant \(\alpha\). Note that the setting is more restrictive compared to the previous one. The goal is to get an upper error bound which is polynomially in \(\alpha\) and \(d\). We consider a Metropolis algorithm based on a ball walk. For \(\delta > 0\) the transition kernel of the \(\delta\) ball walk is

\[ B_\delta(x, A) = \frac{\text{vol}_d(A \cap B_\delta(x))}{\text{vol}_d(B_\delta(0))} + 1_A(x) \left( 1 - \frac{\text{vol}_d(G \cap B_\delta(x))}{\text{vol}_d(B_\delta(0))} \right), \quad x \in G, A \in \mathcal{B}(G), \]
where $B_\delta(x)$ denotes the Euclidean ball with radius $\delta$ around $x$. Let $K_{p,\delta}$ be the transition kernel of the Metropolis algorithm with ball walk proposal $B_\delta$, let $P_{p,\delta}$ be the corresponding transition operator and let $\Lambda_{p,\delta}$ be the largest element of the spectrum of $P_{p,\delta} - \bar{S}: L_2(\rho_p) \to L_2(\rho_p)$. 

In [12 Corollary 1] the following result is proven.

**Proposition 3.** Let $\rho \in \mathcal{R}_{a,d}$ and let $\delta = \min\{1/\sqrt{d+1}, \alpha^{-1}\}$. Then, the conductance of $K_{p,\delta}$ is bounded from below by

$$
0.0025 \frac{\sqrt{d+1}}{\min\{1, \frac{1}{\alpha}\}}.
$$

By Proposition 1 and Proposition 3 we have a lower bound of $1 - \Lambda_{p,\delta}$. However, to apply Theorem 1 we need a lower bound on $\mathbb{E}_\rho$.

Let $\mathcal{F}_{a,d} = \{(f,\rho) : \rho \in \mathcal{R}_{a,d}, f \in L_4(\rho_p), \|f\|_4 \leq 1\}$

and recall that $\mathcal{R}_{a,d}$ is defined in (12). Note that we assumed $G = B_d$. Now we can apply Theorem 1 for the lazy Metropolis algorithm with ball walk proposal $\bar{K}_{p,\delta}$.

**Theorem 5.** Let $\nu$ be the uniform distribution on $B_d$ and let us assume that $\delta = \min\{1/\sqrt{d+1}, \alpha^{-1}\}$. Let $(X_n)_{n \in \mathbb{N}}$ be a Markov chain with transition kernel $\bar{K}_{p,\delta}$, i.e. the lazy version of the Metropolis algorithm with ball walk proposal $B_\delta$, and initial distribution $\nu$. Let

$$
n_0 = \lfloor 5.92 \cdot 10^6(d+1) \max\{\alpha^2, d+1\}(2\alpha + 4.16) \rfloor.
$$

Then

$$
\sup_{(f,G) \in \mathcal{F}_{a,d}} e_\nu(S_n, n_0, (f,\rho)) \leq 1089 \frac{\sqrt{d+1} \max\{\alpha, \sqrt{d+1}\}}{\sqrt{n}}
$$

$$
+ 8.38 \cdot 10^{-1} \frac{\sqrt{d+1} \max\{\alpha^2, d+1\}}{n}.
$$

The last theorem states that the number of oracle calls of $f$ and $\rho$ to obtain an error $\varepsilon > 0$ is bounded by $\kappa d \max\{\alpha^2, d\} (\varepsilon^2 + \alpha)$. Hence the computation of $S(f,\rho)$ is polynomially tractable. Note that $\mathcal{R}_{a,d}$ might be interpreted as a subclass of $\mathcal{R}_C$ with $C = \exp(2\alpha)$ and $G = B_d$, since $\rho \in \mathcal{R}_{a,d}$ implies $\exp(2\alpha) \|\rho\|_\infty \in \mathcal{R}_{\exp(2\alpha)}$. 
Thus, by Theorem 5 we obtain that the number of oracle calls to get an error $\varepsilon$ also depends polynomially on $\log C$, since $C = \exp(2\alpha)$.

4 Open problems and related comments

- We do not know whether an error bound as in Theorem 1 holds for $f \in L_2$ if $\text{gap}(P) > 0$.
- In [16] error bounds of $S_{n,n_0}$ for $f \in L_p$ with $1 < p \leq 2$ are proven. Then one needs a new error criterion, here the absolute mean error

$$E_{\nu,K}|S_{n,n_0}(f) - S(f)|$$

is used. If the Markov chain is $L_1$-exponentially convergent, then the error bound decreases with $n^{1/p-1}$. For a Markov chain with $L_2$-spectral gap a similar error bound is shown.
- The tractability results in Theorem 2 and Theorem 5 are nice since the degree of the polynomial is small. Nevertheless, the upper bound is not really useful because of the huge constants. Is it possible to prove these or similar results with much smaller constants?
- A related question would be the construction of Markov chain quasi-Monte Carlo methods, see [2, 3]. Here the idea is to derandomize the Markov chain by using a carefully constructed deterministic sequence of numbers to obtain a sample $x_1, \ldots, x_{n+n_0}$. However, explicit constructions with small error bounds are not known.

References

1. A. Belloni and V. Chernozhukov, On the computational complexity of MCMC-based estimators in large samples, Ann. Statist. 37 (2009), no. 4, 2011–2055.
2. S. Chen, J. Dick, and A. Owen, Consistency of Markov chain quasi-Monte carlo on continuous state spaces, Ann. Statist. 39 (2011), 673–701.
3. J. Dick, D. Rudolf, and H. Zhu, Discrepancy bounds for uniformly ergodic Markov chain quasi-Monte carlo, Preprint, Available at http://arxiv.org/abs/1303.2423 (2013).
4. A. Joulin and Y. Ollivier, Curvature, concentration and error estimates for Markov chain Monte Carlo, Preprint, Available at http://arxiv.org/abs/1303.2423 (2013).
5. K. Łatuszynski, B. Miasojedow, and W. Niemiro, Nonasymptotic bounds on the estimation error of MCMC algorithms, Bernoulli 19 (2013), no. 5A, 2033–2066.
6. G. Lawler and A. Sokal, Bounds on the $L^2$ spectrum for Markov chains and Markov processes: a generalization of Cheeger’s inequality, Trans. Amer. Math. Soc. 309 (1988), no. 2, 557–580.
7. L. Lovász and S. Vempala, Hit-and-run from a corner, SIAM J. Comput. 35 (2006), no. 4, 985–1005.
8. P. Mathé and E. Novak, Simple Monte Carlo and the Metropolis algorithm, J. Complexity 23 (2007), no. 4-6, 673–696.
9. K. Mengersen and R. Tweedie, *Rates of convergence of the Hastings and Metropolis algorithms*, Ann. Statist. 24 (1996), no. 1, 101–121.

10. E. Novak and H. Woźniakowski, *Tractability of multivariate problems. Vol. 1: Linear information*, EMS Tracts in Mathematics, vol. 6, European Mathematical Society (EMS), Zürich, 2008.

11. E. Novak and H. Woźniakowski, *Tractability of multivariate problems. Vol. 2: Standard information for functionals*, EMS Tracts in Mathematics, vol. 12, European Mathematical Society (EMS), Zürich, 2010.

12. E. Novak and H. Woźniakowski, *Tractability of multivariate problems. Vol. 3: Standard information for operators*, EMS Tracts in Mathematics, vol. 18, European Mathematical Society (EMS), Zürich, 2012.

13. G. Roberts and J. Rosenthal, *Geometric ergodicity and hybrid Markov chains*, Electron. Comm. Probab. 2 (1997), no. 2, 13–25.

14. G. Roberts and J. Rosenthal, *General state space Markov chains and MCMC algorithms*, Probability Surveys 1 (2004), 20–71.

15. G. Roberts and R. Tweedie, *Geometric $L^2$ and $L^1$ convergence are equivalent for reversible Markov chains*, J. Appl. Probab. 38A (2001), 37–41.

16. D. Rudolf and N. Schweizer, *Error bounds of MCMC for functions with unbounded stationary variance*, Preprint, Available at http://arxiv.org/abs/1312.4344 (2013).

17. D. Rudolf and M. Ullrich, *Positivity of hit-and-run and related algorithms*, Electron. Commun. Probab. 18 (2013), 1–8.

18. D. Rudolf, *Hit-and-run for numerical integration*, To appear in: J. Dick, F. Y. Kuo, G. Peters, I. H. Sloan (Eds.), Monte Carlo and Quasi-Monte Carlo Methods 2012, Springer-Verlag.

19. D. Rudolf, *Explicit error bounds for Markov chain Monte Carlo*, Dissertationes Math. 485 (2012), 93 pp.

20. R. Smith, *Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions*, Oper. Res. 32 (1984), no. 6, 1296–1308.