Estimating the spectral gap of a trace-class Markov operator

Qian Qin
Joint work with James P. Hobert and Kshitij Khare.

July 2017
Markov chain Monte Carlo (MCMC) is used to estimate multi-dimensional integrals that represent expectations with respect to intractable probability distributions. Let $\pi$ be an intractable pdf and let

$$J = \int_S f(u)\pi(u) \mu(du).$$

One can simulate a Markov chain $\Phi = \{\Phi_k\}_{k=0}^{\infty}$ that converges to $\pi$ and estimate $J$ by $J_m = m^{-1} \sum_{k=0}^{m-1} f(\Phi_k)$. 
Introduction

Given $f$, the accuracy of the estimation essentially depends on two factors.

1. The convergence rate of $\Phi$, and
2. The correlation between the $f(\Phi_k)$s under stationarity.

These two factors can be investigated jointly under an operator theory framework. They are largely dependent on the spectrum and in particular, the spectral gap of the Markov operator associated with $\Phi$. 
Let $P$ be the Markov operator associated with $\Phi$.

Denote the spectral gap of $P$ by $\delta$. Then $0 \leq \delta \leq 1$. Suppose $\Phi$ is reversible, then

1. 
\[ d_{TV}(\Phi_k; \Phi_\infty) \leq C(1 - \delta)^k, \]
where $d_{TV}(\Phi_k; \Phi_\infty)$ is the total variation distance between the distribution of $\Phi_k$ and the stationary distribution of $\Phi$.

2. Moreover, $(1 - \delta)^k$ is the maximum absolute correlation between $\Phi_j$ and $\Phi_{j+k}$ as $j \to \infty$. This implies that
\[
\limsup_{m \to \infty} \text{var}\left[ m^{1/2} (J_m - J) \right] \leq \frac{2 - \delta}{\delta} \text{var}_{\pi} f.
\]

Goal: Estimate $\delta$. 

Estimating (bounding) $\delta$


Computational approach: Finite-rank approximation, random matrix approximation (Koltchinskii and Giné, 2000).

Simulation approach: autocorrelation plot and others (Garren and Smith, 2000).
Markov operators

$(S, \mathcal{U}, \mu)$ is a countably generated, $\sigma$-finite measure space.

Define a (separable) Hilbert space consisting of complex valued functions on $S$ that are square integrable with respect to $\pi(u)$, namely

$$L^2(\pi) := \left\{ f : S \rightarrow \mathbb{C} \mid \int_S |f(u)|^2 \pi(u) \mu(du) < \infty \right\}.$$

For $f, g \in L^2(\pi)$, their inner product is given by

$$\langle f, g \rangle_\pi = \int_S f(u)\overline{g(u)}\pi(u) \mu(du).$$
Markov operators

Let $p(u, u')$, $u, u' \in S$ be the Markov transition density (Mtd) that gives rise to $\Phi$, i.e. for any $A \in \mathcal{U}$

$$
\mathbb{P}(\Phi_k \in A|\Phi_0 = u) = \int_A p^{(k)}(u, u') \mu(du'),
$$

where

$$
p^{(k)}(u, u') := \begin{cases} p(u, u') & k = 1, \\
\int_S p^{(k-1)}(u, w)p(w, u') \mu(dw) & k > 1.
\end{cases}
$$

The transition density $p(u, u')$ defines the following linear (Markov) operator $P$. For any $f \in L^2(\pi)$,

$$
Pf(u) = \int_S p(u, u')f(u') \mu(du').
$$
We say that $P$ is trace-class if it is compact and has absolutely summable eigenvalues.

Suppose $P$ is non-negative and trace-class. Then all the eigenvalues of $P$ are non-negative. Let $\{\lambda_i\}_{i=0}^{\infty}$ be the (positive) eigenvalues of $P$ in decreasing order, taking into account multiplicity. Then $\lambda_0 = 1$, and $\sum_{i=0}^{\infty} \lambda_i < \infty$. Under mild assumptions, we have $\lambda_1 < 1$.

The spectral gap $\delta = 1 - \lambda_1$, where $\lambda_1$ is the second largest eigenvalue of $P$.

Question: How to estimate $\lambda_1$?
Power sums of eigenvalues

For \( k \in \mathbb{N} \), let \( s_k = \sum_{i=0}^{\infty} \lambda_i^k \). Let \( u_k = (s_k - 1)^{1/k} \) and
\( l_k = (s_k - 1)/(s_{k-1} - 1) \). Then we have the following.

**Proposition**

As \( k \to \infty \),
\[
\begin{align*}
u_k & \downarrow \lambda_1, \\
l_k & \uparrow \lambda_1.
\end{align*}
\]

To bound \( \lambda_1 \), we can consider estimating the \( s_k \)s. We will make use of the following trace formula
\[
s_k = \int_S p^{(k)}(u, u) \mu(du).
\]
Data augmentation (DA) operators

Let $S_U = S$ and $\pi_U(u) = \pi(u)$. Define $(S_V, \mathcal{V}, \nu)$ to be a $\sigma$-finite measure space such that $\mathcal{V}$ is countably generated. Consider the random element $(U, V)$ taking values in $S_U \times S_V$ with joint pdf $\pi_{U,V}(u, v)$. Suppose that the marginal pdf of $U$ is $\pi_U(u)$ and denote the marginal pdf of $V$ by $\pi_V(v)$.

We call $\Phi$ a DA chain, and accordingly, $P$ a DA operator, if $p(u, u')$ can be expressed as

$$p(u, u') = \int_{S_V} \pi_{U|V}(u' | v) \pi_{V|U}(v | u) \, \nu(dv).$$

This chain is reversible with respect to $\pi_U := \pi$. 
Data augmentation (DA) operators

Mtd:
\[ p(u, u') = \int_{S_v} \pi_{U|V}(u'|v) \pi_{V|U}(v|u) \nu(dv). \]

To simulate a DA chain, we need to be able to sample from \( \pi_{U|V}(\cdot|v) \) and \( \pi_{V|U}(v|u) \). Simulation process: \( u \rightarrow v \rightarrow u' \). Here, \( v \) is a latent variable. Alternatively, one can simply view DA as the marginal chain of a Gibbs sampler.

A DA operator is necessarily non-negative.

Note that even if \( \Phi \) is reversible but not a DA chain, \( \{\Phi_{2k}\}_{k=0}^{\infty} \) is. Note that the corresponding Mtd is
\[ p^{(2)}(u, u') = \int_S p(u, v)p(v, u') \mu(dv). \]

If we take
\[ \pi_{U,V}(u, v) = \pi(u)p(u, v) = \pi(v)p(v, u) \]
then \( \pi_{U|V}(u'|v) = p(v, u') \), and \( \pi_{V|U}(v|u) = p(v, u) \).
Integral representation of $s_k$

Theorem

The DA operator $P$ is trace-class if and only if

$$\int_{S_U} p(u, u) \mu(du) := \int_{S_U} \int_{S_V} \pi_{U|V}(u|v) \pi_{V|U}(v|u) \nu(dv) \mu(du) < \infty. \quad (1)$$

If (1) holds, then for any positive integer $k$,

$$s_k := \sum_{i=0}^{\infty} \lambda_i^k = \int_{S_U} p^{(k)}(u, u) \mu(du).$$

In order to find $s_k$, $k \in \mathbb{N}$, all we need is to evaluate $\int_{S_U} p^{(k)}(u, u) \mu(du)$. This is in general not easy. We will introduce a way of estimating these integrals using classical Monte Carlo.
Estimating $s_k$

Let $\psi : S_U \rightarrow (0, \infty)$ be a pdf that's positive everywhere. Then

$$
\int_{S_U} p^{(k)}(u, u) \mu(du) 
= \int_{S_V} \int_{S_U} \frac{\pi_{U|V}(u|v)}{\psi(u)} 
\times \left( \int_{S_U} \pi_{V|U}(v|w) p^{(k-1)}(u, w) \mu(dw) \right) \psi(u) \mu(du) \nu(dv).
$$

Note that

$$
\eta(u, v) := \left( \int_{S_U} \pi_{V|U}(v|w) p^{(k-1)}(u, w) \mu(dw) \right) \psi(u)
$$

is a pdf on $S_U \times S_V$. 
Estimating $s_k$

Recall that

$$s_k = \int_{S_U} p^{(k)}(u, u) = \int_{S_V} \int_{S_U} \frac{\pi_{U|V}(u|v)}{\psi(u)} \eta(u, v) \mu(du) \nu(dv),$$

where

$$\eta(u, v) := \left( \int_{S_U} \pi_{V|U}(v|w) p^{(k-1)}(u, w) \mu(dw) \right) \psi(u).$$

Suppose that \(\{U^*, V^*\} \sim \eta\). Then

$$s_k = \mathbb{E} \frac{\pi_{U|V}(U^*|V^*)}{\psi(U^*)} \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\pi_{U|V}(U_i^*|V_i^*)}{\psi(U_i^*)},$$

where \(\{U_i^*, V_i^*\}_{i=1}^{N}\) are iid copies of \((U^*, V^*)\).
Estimating $s_k$

How to simulate $\eta$? Recall that

$$\eta(u, v) := \left( \int_{S_U} \pi_{V|U}(v|w)p^{(k-1)}(u, w) \mu(dw) \right) \psi(u).$$

One can use the algorithm below.

---

Algorithm 1: $i$th iteration. $(U^*, V^*) \sim \eta$

1. Generate $U^*$ from $\psi(u)$.
2. If $k = 1$, set $W = U^*$. If $k \geq 2$, given $U^* = u$, generate $W$ from $p^{(k-1)}(u, w)$ by running $k - 1$ iterations of the DA algorithm of interest.
3. Given $W = w$, generate $V^*$ from $\pi_{V|U}(v|w)$.

---
Estimating $s_k$

For the estimation to be statistically valid, we’d like the estimator to have finite variance, i.e.

$$D^2 := \text{var} \left( \frac{\pi_{U\mid V}(U^* \mid V^*)}{\psi(U^*)} \right) < \infty.$$ 

The following theorem provides a sufficient condition for this to be true.

**Theorem**

The variance, $D^2$, is finite if

$$\int_{S_V} \int_{S_U} \pi^3_{U\mid V}(u\mid v) \pi_{V\mid U}(v\mid u) \frac{\mu(du) \nu(dv)}{\psi^2(u)} < \infty.$$
Estimating $s_k$

Recall that

$$D^2 := \text{var} \left( \frac{\pi_{U|V}(U^*|V^*)}{\psi(U^*)} \right)$$

$$= \int_{S_U \times S_V} \frac{\pi_{U|V}^2(u|v)}{\psi^2(u)} \left( \int_{S_U} \pi_{V|U}(v|w)p^{(k-1)}(u, w) \mu(dw) \right) \psi(u) dvdu - s_k^2.$$  

(Variance of the estimator is $D^2 / N$.)

Heuristically, if $\psi \approx \pi_U$, then as $k \to \infty$,

$$D^2 \approx \int_{S_U \times S_V} \frac{\pi_{U|V}^2(u|v)}{\pi_U^2(u)} \pi_V(v) dvdu - 1,$$

i.e. $D^2 \approx s_1 - 1$. Therefore, it’s beneficial to choose $\psi$ that resembles the target distribution if the sum of eigenvalues, $s_1$, is small.
Let \( S_U = S_V = \mathbb{R} \), \( \pi_U(u) \propto \exp(-u^2) \), and
\[
\pi_{V|U}(v|u) \propto \exp\left\{ -4 \left( v - \frac{u}{2} \right)^2 \right\}.
\]

Then
\[
\pi_{U|V}(u|v) \propto \exp\{-2(u - v)^2\}.
\]

This characterizes one of the simplest DA chains known, with Mtd
\[
p(u, u') = \int_{\mathbb{R}} \pi_{U|V}(u'|v)\pi_{V|U}(v|u) \, dv
\]

being the pdf of a normal distribution.

The spectrum of the corresponding Markov operator \( P \) has been studied thoroughly. It’s easy to verify that \( P \) is trace-class. In fact, for any non-negative integer \( i \), \( \lambda_i = 1/2^i \). This implies for any positive integer \( k \),
\[
S_k = \sum_{i=0}^{\infty} \frac{1}{2^{ik}} = \frac{1}{1 - 2^{-k}}.
\]
Illustration

With $N = 10^5$, our estimates for $s_k$, $k = 1, 2, 3, 4$ are as follows.

**Table:** Estimated power sums of eigenvalues for the Gaussian chain

<table>
<thead>
<tr>
<th>$k$</th>
<th>True $s_k$</th>
<th>Est. $s_k$</th>
<th>Est. $D/\sqrt{N}$</th>
<th>Est. $l_k$</th>
<th>Est. $u_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.000</td>
<td>1.996</td>
<td>0.004</td>
<td>0.000</td>
<td>0.996</td>
</tr>
<tr>
<td>2</td>
<td>1.333</td>
<td>1.331</td>
<td>0.004</td>
<td>0.333</td>
<td>0.575</td>
</tr>
<tr>
<td>3</td>
<td>1.143</td>
<td>1.142</td>
<td>0.004</td>
<td>0.429</td>
<td>0.522</td>
</tr>
<tr>
<td>4</td>
<td>1.067</td>
<td>1.068</td>
<td>0.004</td>
<td>0.482</td>
<td>0.511</td>
</tr>
</tbody>
</table>
Let $Y_1, Y_2, \ldots, Y_n$ be independent Bernoulli random variables with $P(Y_i = 1 | \beta) = \Phi(x_i^T \beta)$, where $x_i, \beta \in \mathbb{R}^p$. Take the prior on $\beta$ to be $N_p(Q^{-1} \nu, Q^{-1})$, where $\nu \in \mathbb{R}^p$ and $Q$ is positive definite. The resulting posterior distribution is intractable, but Albert and Chib (1993) devised a DA algorithm to sample from it.

Posterior:

\[
\pi(\beta | Y) \propto \prod_{i=1}^{n} \left( \Phi(x_i^T \beta) \right)^{y_i} \left( 1 - \Phi(x_i^T \beta) \right)^{1-y_i} \exp \left\{ -\frac{1}{2} (\beta - Q^{-1} \nu)^T Q (\beta - Q^{-1} \nu) \right\}.
\]

Albert and Chib’s chain:

\[
\begin{align*}
Z_i | \beta & \sim \begin{cases} 
TN(x_i^T \beta, 0, \infty), & Y_i = 1, \\
TN(x_i^T \beta, -\infty, 0), & Y_i = 0;
\end{cases} \\
\beta | z & \sim N \left( (X^T X + Q)^{-1} (X^T z + \nu), (X^T X + Q)^{-1} \right).
\end{align*}
\]
Chakraborty and Khare (2017) showed that when all the eigenvalues of $Q^{-1/2}X^TXQ^{-1/2}$ are less than $7/2$, then the corresponding Markov operator is trace-class.

We will use our method to estimate the spectral gap of the chain. The dataset we examine is the "lupus" data (van Dyk 2001), which has $n = 55$ observations and $p = 3$ features.

For the prior, we take $\nu = 0$, and $Q = X^TX/3.499999$. This is a $g$-prior-like prior that Chakraborty and Khare used.
$N = 4 \times 10^5$.

**Table:** Estimated power sums of eigenvalues for the AC chain

<table>
<thead>
<tr>
<th>$k$</th>
<th>Est. $s_k$</th>
<th>Est. $D/\sqrt{N}$</th>
<th>Est. $l_k$</th>
<th>Est. $u_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.744</td>
<td>0.072</td>
<td>0.000</td>
<td>5.744</td>
</tr>
<tr>
<td>2</td>
<td>2.041</td>
<td>0.007</td>
<td>0.181</td>
<td>1.020</td>
</tr>
<tr>
<td>3</td>
<td>1.363</td>
<td>0.004</td>
<td>0.349</td>
<td>0.713</td>
</tr>
<tr>
<td>4</td>
<td>1.156</td>
<td>0.004</td>
<td>0.430</td>
<td>0.628</td>
</tr>
<tr>
<td>5</td>
<td>1.068</td>
<td>0.003</td>
<td>0.436</td>
<td>0.584</td>
</tr>
</tbody>
</table>

By CLT, a (conservative) asymptotic 95% CI for $\lambda_1$ is $(0.397, 0.595)$. 


