Estimating the spectral gap of a trace-class Markov operator

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

July 2017

Introduction

Markov chain Monte Carlo (MCMC) is used to estimate multi-dimensional integrals that represent expectations with respect to intractable probability distributions. Let π 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 π 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 Φ, 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 Φ.

Introduction

Let P be the Markov operator associated with Φ.

Denote the spectral gap of P by δ. Then $0 \leq \delta \leq 1$. Suppose Φ 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 Φ_k and the stationary distribution of Φ.
2. Moreover, $(1-\delta)^k$ is the maximum absolute correlation between Φ_j and Φ_{j+k} as $j \to \infty$. This implies that

   $$\limsup_{m\to\infty} \mathrm{var}\left[m^{1/2}(J_m - J)\right] \leq \frac{2-\delta}{\delta}\mathrm{var}_\pi f.$$

Goal: Estimate δ.

Estimating (bounding) δ

Theoretical approach: Path arguments (Diaconis and Stroock, 1991), conductance and Cheeger's inequality (Lawler and Sokal, 1988; Sinclair and Jerrum, 1989), drift and minorization (Rosenthal, 1995).

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, σ-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

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

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 Φ, 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 $f \in L^2(\pi)$,

$$Pf(u) = \int_S p(u, u') f(u')\, \mu(du').$$

Markov operators

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 λ_1 is the second largest eigenvalue of P.

Question: How to estimate λ_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 $k \to \infty$,

$$u_k \downarrow \lambda_1,$$
$$l_k \uparrow \lambda_1.$$

To bound λ_1, we can consider estimating the s_ks. 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 σ-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 Φ 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: $u \to v \to u'$. 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 Φ 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

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

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

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

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 η? 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.

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

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

Theorem

The variance, D^2, is finite if

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

Estimating s_k

Recall that

$$
\begin{aligned}
&D^2 \\
&:= \mathrm{var}\left(\frac{\pi_{U|V}(U^*|V^*)}{\psi(U^*)}\right) \\
&= \int_{S_U \times S_V} \frac{\pi_{U|V}^2(u|v)}{\psi^2(u)} \left(\int_{S_U} \pi_{V|U}(v|w) p^{(k-1)}(u,w)\, \mu(dw)\right) \psi(u)\, dvdu - s_k^2.
\end{aligned}
$$

(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 ψ that resembles the target distribution if the sum of eigenvalues, s_1, is small.

Illustration

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

k	True s_k	Est. s_k	Est. $D/\sqrt{N}$	Est. l_k	Est. u_k
1	2.000	1.996	0.004	0.000	0.996
2	1.333	1.331	0.004	0.333	0.575
3	1.143	1.142	0.004	0.429	0.522
4	1.067	1.068	0.004	0.482	0.511

Illustration

Let $Y_1, Y_2, \ldots, Y_n$ be independent Bernoulli random variables with $\mathbb{P}(Y_i = 1|\beta) = \Phi(x_i^T \beta)$, where $x_i, \beta \in \mathbb{R}^p$. Take the prior on β to be $N_p(Q^{-1}v, Q^{-1})$, where $v \in \mathbb{R}^p$ 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:

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

Albert and Chib's chain:

$$z_i|\beta \sim \begin{cases} TN(x_i^T \beta, 0, \infty), & Y_i = 1, \\ TN(x_i^T \beta, -\infty, 0), & Y_i = 0; \end{cases}$$
$$\beta|z \sim N\left((X^T X + Q)^{-1}(X^T z + v), (X^T X + Q)^{-1}\right).$$

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 $v = 0$, and $Q = X^TX/3.499999$. This is a g-prior-like prior that Chakraborty and Khare used.

Illustration

$N = 4 \times 10^5$.

Table: Estimated power sums of eigenvalues for the AC chain

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

By CLT, a (conservative) asymptotic 95% CI for λ_1 is $(0.397, 0.595)$.

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88** 669–679.

CHAKRABORTY, S. and KHARE, K. (2017). Convergence properties of Gibbs samplers for Bayesian probit regression with proper priors. *Electronic Journal of Statistics* **11** 177–210.

DIACONIS, P. and STROOCK, D. (1991). Geometric bounds for eigenvalues of Markov chains. *The Annals of Applied Probability* **1** 36–61.

GARREN, S. T. and SMITH, R. L. (2000). Estimating the second largest eigenvalue of a Markov transition matrix. *Bernoulli* **6** 215–242.

KOLTCHINSKII, V. and GINÉ, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* **6** 113–167.

LAWLER, G. F. and SOKAL, A. D. (1988). Bounds on the l^2 spectrum for Markov chains and Markov processes: A generalization of Cheegers inequality. *Transactions of the American Mathematical Society* **309** 557–580.

ROSENTHAL, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association* **90** 558–566.

SINCLAIR, A. and JERRUM, M. (1989). Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation* **82** 93–133.