Spectral properties of Markov operators
in Markov chain Monte Carlo

Qian Qin
Advisor: James P. Hobert
October 2017

1 Introduction

Markov chain Monte Carlo (MCMC) is an indispensable tool in Bayesian statistics. Consider an intractable probability measure on a space, $\mathcal{X}$, that we denote by $\Pi(\cdot)$. In a Bayesian setting, $\Pi$ is usually the posterior distribution of some multi-dimensional parameter that one wants to estimate. Researchers would be interested in finding the expectations of functions of these parameters with respect to $\Pi$, in other words, quantities of the form

$$E_{\Pi}f = \int_{\mathcal{X}} f(x)\Pi(dx).$$

$\Pi$ being intractable means that the above integral cannot be evaluated analytically. Moreover, due to the multi-dimensionality of $\mathcal{X}$, numerical methods would often be impractical. The integral can be estimated using classical Monte Carlo, if one can sample from $\Pi$. However, it is rarely the case that $\Pi$ is standard, and direct sampling is usually highly inefficient. It is under these circumstances that MCMC comes into play. To estimate $E_{\Pi}f$ using an MCMC algorithm, one simulates a Markov chain, $\{X_k\}_{k=0}^{\infty}$ on $\mathcal{X}$ ($\mathcal{X}$ will be called the chain’s state space), such that $X_k \xrightarrow{d} X$, where $X \sim \Pi$. One then uses

$$\bar{X}_m = m^{-1} \sum_{k=0}^{m-1} f(X_k)$$

as an estimate for $E_{\Pi}f$.

It’s evident that the accuracy of the above estimation depends largely on the rate that $X_k$ converges to $X \sim \Pi$. This rate is controlled by the spectrum of a certain linear operator associated with the chain, which people refer to as the Markov operator. In what follows, I will try to present a gentle introduction to Markov operators and their spectra, and how they are studied by MCMC theorists.
2 Transition functions and matrices

A Markov chain is a sequence of random elements such that given the \((k - 1)\)th element, the distribution of the \(k\)th element is independent of all elements preceding the \((k - 1)\)th one. In other words, for any measurable set \(A \subset \mathcal{X}\),

\[
\mathbb{P}(X_k \in A | X_i = x_i, i = 1, 2, \ldots, k - 1) = \mathbb{P}(X_k \in A | X_{k-1} = x_{k-1}).
\]

We will be studying only time-homogeneous chains, that is, chains such that the distribution of \(X_k | X_{k-1} = x\) depends only on \(x\), but not on \(k\). In such cases, for each \(x \in \mathcal{X}\), the distribution of \(X_k | X_{k-1} = x\) is characterized by a transition measure \(P(x, \cdot)\), such that

\[
\mathbb{P}(X_1 \in A | X_0 = x) = \mathbb{P}(X_k \in A | X_{k-1} = x)
\]

for any measurable \(A\). Then \(P(\cdot, \cdot)\) is called the Markov transition function (MtF) of the chain. It’s often the case that, for each \(x \in \mathcal{X}\), \(P(x, \cdot)\) is dominated by some underlying measure, so that

\[
P(x, dx') = p(x, x') dx',
\]

where \(p(x, \cdot)\) is a probability density function (pdf) on \(\mathcal{X}\). In such cases, \(p(\cdot, \cdot)\) is referred to as the Markov transition density (MtD) of the chain.

Consider a particularly simple scenario, where the state space, \(\mathcal{X}\), is discrete and finite, say, \(\mathcal{X} = \{1, 2, \ldots, r\}\). A Markov chain on the state space is then associated with a MtD with respect to the counting measure:

\[
\mathbb{P}(X_k = j | X_{k-1} = i) = \mathbb{P}(X_1 = j | X_0 = i) = p(i, j), \ (i, j) \in \mathcal{X}^2.
\]

We can arrange \(p(\cdot, \cdot)\) into an \(r \times r\) matrix \(P\), such that its \(ij\)th element is \(p(i, j)\). This matrix is then called the chain’s Markov transition matrix, or MtM. The matrix is stochastic, in the sense that each of its row is a probability vector, i.e., for each \(i \in \{1, 2, \ldots, r\}\), \(p(i, j) \geq 0\), and \(\sum_{j=1}^{r} p(i, j) = 1\). Although we are almost never interested in discrete state spaces in Bayesian statistics, we will investigate them first, and draw analogies to them when dealing with more complicated setups. For the rest of this section, assume that \(\mathcal{X} = \{1, 2, \ldots, r\}\).

Basic arithmetics can be applied to transition matrices, and some of them have natural interpretations. For example, multiplying \(P\) by itself \(k - 1\) times yields \(P^k\), which is also an \(r \times r\) stochastic matrix. It’s easy to see that its \(ij\)th element is

\[
p^{(k)}(i, j) = \mathbb{P}(X_k = j | X_0 = i) = \mathbb{P}(X_{k+k'} = j | X_{k'} = i), \ k' \in \mathbb{Z}_+.
\]
Suppose that the Markov chain corresponding to \( p(\cdot, \cdot) \) converges to a stationary distribution, \( \Pi \), that is \( X_k \xrightarrow{d} X \sim \Pi \) regardless of how \( X_0 \) is distributed. Then for each \( i \in \mathcal{X} \),

\[
\lim_{k \to \infty} \sum_{j=1}^{r} |p^{(k)}(i, j) - \pi(j)| = 0, \tag{1}
\]

where \( \pi(\cdot) \) is the pdf of \( \Pi \). In the above equation,

\[
\sum_{j=1}^{r} |p^{(k)}(i, j) - \pi(j)| = 2 \sup_{A \subseteq \mathcal{X}} |P(X_k \in A | X_0 = i) - \Pi(A)|
\]

is called the total variation distance between the distribution of \( X_k | X_0 = i \) and the chain’s stationary distribution, \( \Pi \). Note that if (1) holds, then applying the dominated convergence theorem, one can see that for any \( j \in \mathcal{X} \),

\[
\pi(j) = \lim_{k \to \infty} \sum_{j'=1}^{r} p^{(k)}(i, j')p(j', j) = \sum_{j'=1}^{r} \pi(j')p(j', j). \tag{2}
\]

It’s in this sense that we call \( \Pi \) “stationary”.

One can also multiply \( P \) to a vector in \( \mathbb{C}^r \). Let \( f = (f(1), f(2), \ldots, f(r))^T \in \mathbb{C}^r \). Then \( f \) can viewed as a function on \( \mathcal{X} = \{1, 2, \ldots, r\} \) that maps \( i \in \mathcal{X} \) to \( f(i) \). \( Pf \) is also a vector in \( \mathbb{C}^r \), and thus can be viewed as a function as well. Its \( i \)th element is

\[
(Pf)(i) = \sum_{j=1}^{r} p(i, j)f(j) = \mathbb{E}(f(X_1) | X_0 = i).
\]

Similarly, \( P^k f \) is a vector whose \( i \)th element is \( \mathbb{E}(f(X_k) | X_0 = i) \). Like all square matrices, \( P \) defines a linear transformation on \( \mathbb{C}^r \). Hence, we can view \( P \) as a linear operator.

A complex number, \( \lambda \), is called an eigenvalue of \( P \) if \( P - \lambda I \) is singular, where \( I \) is the \( r \times r \) identity matrix. For each eigenvalue of \( P \), \( \lambda \), there necessarily exists an eigenvector, \( f : \mathcal{X} \to \mathbb{C} \) such that \( Pf = \lambda f \). Because vectors in \( \mathbb{C}^r \) can also be regarded as functions on the state space, we will sometimes refer eigenvectors as eigenfunctions. Since \( P \) is a stochastic matrix, the vector \( f_0 = (1, 1, \ldots, 1)^T \in \mathbb{C}^r \) is an eigenvector of \( P \), and the associated eigenvalue is \( \lambda_0 = 1 \). The following proposition is well-known (see, e.g., Kontoyiannis and Meyn, 2012).

**Proposition 1.** Suppose that (1) holds for all \( i \in \mathcal{X} \). Then all the eigenvalues of \( P \), excluding \( \lambda_0 = 1 \), are less than 1 in modulus. Moreover, the algebraic multiplicity of \( \lambda_0 \) is 1.

**Proof.** For \( f = (f(1), f(2), \ldots, f(r))^T \in \mathbb{C}^r \), let \( \mathbb{E}_\Pi f = \sum_{i=1}^{r} f(i)\pi(i) \) be the expectation of \( f \) under \( \Pi \). It follows from (1) that for any and any \( i \in \mathcal{X} \),

\[
\lim_{k \to \infty} |(P^k f)(i) - \mathbb{E}_\Pi f| \leq \lim_{k \to \infty} \sum_{j=1}^{r} |p^{(k)}(i, j) - \pi(j)| \max_{j'} |f(j')| = 0. \tag{3}
\]
Suppose that there exists an eigenvalue $\lambda \neq 1$ such that $|\lambda| \geq 1$. Then $Pf = \lambda f$ for some $f \in \mathcal{X} \to \mathbb{C}$ that’s not proportional to $f_0 = (1, 1, \ldots, 1)^T$. Thus, $(P^k f)(i) = \lambda^k f(i)$ for all $i \in \mathcal{X}$ and $k \geq 1$. Clearly, (3) cannot hold for all $i \in \mathcal{X}$. Therefore, such an eigenvalue does not exist. Similarly, one can show that there doesn’t exist a function $f : \mathcal{X} \to \mathbb{C}$ such that $f$ isn’t proportional to $f_0$, and that $Pf = f$. This is to say, the geometric multiplicity of $\lambda_0$ is 1. Finally, suppose that the algebraic multiplicity of $\lambda_0$ is greater than 1, then there exists $f : \mathcal{X} \to \mathbb{C}$ such that $Pf = f + f_0$. But then $(P^k f)(i)$ would be unbounded for each $i$ as $k \to \infty$. Hence, such $f$ cannot exist.

As we’ve mentioned, the convergence rate of a Markov chain is particularly important in MCMC. When the state space is finite, the total variation distance between the distribution of $X_k|X_0 = i$ and the stationary distribution is controlled by the spectrum of $P$. Denote by $\lambda_1$ the second largest eigenvalue of $P$. Then we have the following classical result.

**Proposition 2.** Suppose that (2) holds. Suppose further that for $P$, the algebraic multiplicity of $\lambda_0 = 1$ is 1, and that there are no other eigenvalues of modulus 1. Then for any $\rho > |\lambda_1|$, there exists a function $M : \mathcal{X} \to [0, \infty)$ such that

$$\sum_{j=1}^r |p^{(k)}(i,j) - \pi(j)| \leq M(i) \rho^k.$$  

To prove the result, we will use the classical spectral radius formula, which we now state.

**Proposition 3.** (Spectral radius formula) Let $K$ be an an $r \times r$ matrix, and let $\lambda_0(K)$ be its largest eigenvalue in modulus. Let $\| \cdot \|$ be an arbitrary matrix norm, then

$$|\lambda_0(K)| = \lim_{k \to \infty} \|K^k\|^{1/k}.$$  

Note that this formula can be extended to linear operators on Banach spaces (see, e.g., Arveson, 2006).

Let $P_0$ be the $r \times r$ matrix whose $ij$th element is $\pi(j)$. Then for any $f : \mathcal{X} \to \mathbb{C},$

$$P_0 f(i) = \sum_{j=1}^r f(j) \pi(j) = \mathbb{E} \Pi f.$$  

Clearly, $P_0$ is a projection onto the subspace spanned by $f_0$. We will use the following lemma.

**Lemma 4.** Suppose that (2) holds. Suppose further that for $P$, and the algebraic multiplicity of $\lambda_0 = 1$ is 1. Then the second largest eigenvalue of $P$ in modulus, $\lambda_1$, is the largest eigenvalue of $P - P_0$ in modulus.

**Proof.** By (2), $P_0$ commutes with $P$. Let $\lambda$ be any eigenvalue of $P$ that’s not $\lambda_0 = 1$, and let $f : \mathcal{X} \to \mathbb{C}$ be a generalized eigenvalue associated with it. Then there exists some $k$ such that $(P - \lambda I)^k f = 0$. Applying
Proposition 2: For $P : \mathcal{X} \to \mathbb{C}$, let $\|f\|_\infty = \max_{1 \leq i \leq r} |f(i)|$, and for any $r \times r$ matrix $K$, let

$$
\|K\|_\infty = \sup_{f \in \mathbb{C}^r, f \neq 0} \frac{\|Kf\|_\infty}{\|f\|_\infty}.
$$

Then $\cdot \|_\infty$ is a proper vector (matrix) norm. By Proposition 3 and Lemma 4, there exists an integer $k_0$ such that when $k \geq k_0$, $\|(P - P_0)^k\|_\infty \leq \rho^k$.

Fix $k \geq k_0$. By (2), for any $f : \mathcal{X} \to \mathbb{C}$ and $i \in \mathcal{X}$,

$$
\sum_{j=1}^{r} \left(p^{(k)}(i, j) - \pi(j)\right) f(j) = ((P^k - P_0)f)(i) = ((P - P_0)^k f)(i),
$$

the $i$th element of $(P - P_0)^k f$. Hence,

$$
\max_{1 \leq i \leq r} \sum_{j=1}^{r} \left(p^{(k)}(i, j) - \pi(j)\right) f(j) \leq \|(P - P_0)^k\|_\infty \|f\|_\infty \leq \rho^k \|f\|_\infty.
$$

For each $i \in \mathcal{X}$ and $k \in \mathbb{Z}_+$, it’s easy to verify that

$$
\sum_{j=1}^{r} |p^{(k)}(i, j) - \pi(j)| = \sup_{\|f\|_\infty = 1} \sum_{j=1}^{r} \left(p^{(k)}(i, j) - \pi(j)\right) f(j).
$$

(Just consider separately the cases that $p^{(k)}(i, j) - \pi(j)$ is positive and negative.) Therefore, whenever $k \geq k_0$,

$$
\max_{1 \leq i \leq r} \sum_{j=1}^{r} |p^{(k)}(i, j) - \pi(j)| \leq \rho^k,
$$

and the result follows. \[\square\]

Suppose that we can find the spectrum of a transition matrix, then we can bound the convergence rate of the corresponding Markov chain. In practice, the state spaces that statisticians care about are usually continuous. In such cases, there are no longer transition matrices available. However, many of the above results can be generalized to suit continuous state spaces. It’s in the next section that we will discuss the generalized version of transition matrices: Markov operators.
3 Markov operators

We now return to a general state space. Since \(X\) is no longer required to be finite or discrete, there is no general way to define a transition matrix for the Markov chain in question. Instead, we should consider the Mtf, or Mtd. Recall that \(P(x, \cdot)\) is the distribution rule of \(X_1\) given that \(X_0 = x\). Similarly, define \(P^k(x, \cdot)\) to be the distribution rule of \(X_k\) given \(X_0 = x\). The stationary distribution, \(\Pi\), is the measure such that

\[
\int_X P(x, \cdot) \Pi(dx) = \Pi(\cdot).
\]

Throughout the rest of this report, we assume that such \(\Pi\) exists and is a probability measure, as in all practically useful MCMC algorithms.

For two probability measures on a measurable \(X\), \(\mu\) and \(\nu\), their total variation distance is

\[
\|\mu - \nu\|_{TV} = \|\mu(\cdot) - \nu(\cdot)\|_{TV} = 2 \sup_A (\mu(A) - \nu(A)),
\]

where the supremum is taken over all measurable subsets of \(X\). We say that the chain converges to its stationary distribution if

\[
\lim_{k \to \infty} \|P^k(x, \cdot) - \Pi(\cdot)\|_{TV} = 0
\]

for (almost) all \(x \in X\). We say that the chain is geometrically ergodic if there exists \(\lambda \in [0, 1)\) and \(M : X \to [0, \infty)\) such that

\[
\|P^k(x, \cdot) - \Pi(\cdot)\|_{TV} \leq M(x)\lambda^k
\]

for all \(k \in \mathbb{Z}_+\). A major problem in MCMC theory is finding \(\lambda\) for Markov chains. We have seen in the discrete case how \(\lambda\) is controlled by the spectrum of the Mtm. For chains on general state spaces, we need to generalize transition matrices to Markov operators.

A Markov operator is a linear operator defined using the Mtf. Roughly speaking, for a function \(f : X \to \mathbb{C}\), we can let \(Pf\) be a function such that

\[
(Pf)(x) = \int_X f(x') P(x, dx') = \mathbb{E}(f(X_1)| X_0 = x).
\]

Note how similar this is to \(Pf\) defined in the discrete case. One can check that the transformation \(f \mapsto Pf\) is linear in \(f\). Hence, we can view \(P\) as a linear operator on some Banach space of functions on \(X\), which we denote by \(B\).

Depending on the type of distance between measures one uses, there are multiple function spaces that may come in handy. For studying convergence in total variation, there are two types of function spaces that are commonly considered.
1. A Hilbert space, $L^2(\pi)$, which consists of functions that are square integrable with respect to $\Pi$. For $f, g \in L^2(\pi)$, their inner product is
\[
\langle f, g \rangle = \int_X f(x)g(x)\Pi(dx).
\]
The norm of $f \in L^2(\pi)$ is, of course, $\|f\|_2 = \sqrt{\langle f, f \rangle}$.

2. A Banach space, $L_V$, which consists of measurable functions, $f$, such that
\[
\sup_{x \in X} \frac{|f(x)|}{V(x)} < \infty,
\]
where $V : X \rightarrow [1, \infty)$ is some fixed function. The left hand side of the above inequality is the norm of $f$, which is denoted by $\|f\|_V$.

Note that if $V \equiv 1$, then $\| \cdot \|_V$ (not to be confused with the $L^1$ norm, $\| \cdot \|_1$) is analogous to $\| \cdot \|_\infty$ in the discrete case. Regardless of which space we use, we will always assume that the constant function, $f_0 \equiv 1$ is in $B$. For a two probability measures on $X$, $\mu$ and $\nu$, their discrepancy can be quantified by the following distance:
\[
\|\mu - \nu\| = \sup_{\|f\|=1} \left| \int_X f(x)(\mu(dx) - \nu(dx)) \right|,
\]
where $\|f\|$ being the vector norm of $f \in B$. If $V \equiv 1$, and $B = L_V$, then the above distance is the total variation distance. If $B = L^2(\pi)$, then the above distance is called the $\chi^2$ distance, which dominates total variation distance. Of course, $\|\mu\|$ can be defined analogously for any signed measure $\mu$.

Having defined a Markov operator on a suitable function space, we can talk about its spectrum. In some important cases, the spectrum of a Markov operator dictates the convergence rate of the corresponding Markov chain. For a linear operator, $K : B \rightarrow B$, its norm is induced by the norm of vectors in $B$, that is,
\[
\|K\| = \sup_{f \in B, f \neq 0} \frac{\|Kf\|}{\|f\|} = \sup_{\|f\|=1} \|Kf\|.
\]
The spectrum of $K$, $\sigma(K)$, is defined to be the set of complex numbers, $\lambda$, such that $(K - \lambda I)^{-1}$ doesn’t exist, or is unbounded. We denote by $\rho(K)$ the spectral radius of $K$, which is $\{|\lambda| : \lambda \in \sigma(K)\}$.

Let $P_0$ be a linear operator on $B$ such that for any $f \in B$,
\[
P_0 f(x) = \int_X f(x')\Pi(dx') = \mathbb{E}_\Pi f.
\]
This operator was defined analogously in the discrete case. It maps any function in $B$ to a function proportional to $f_0$. The following (somewhat) well-known result is analogous to Proposition 2.
**Proposition 5.** Suppose that $B$ is a Banach space equipped with the norm $\| \cdot \|$. Then for any $\rho > \rho(P - P_0)$, there exists $k_0 \in \mathbb{Z}_+$ such that whenever $k \geq k_0$,

$$\sup_{f \in B, \| f \| = 1} \| (P^k - P_0)f \| \leq \rho^k.$$ 

**Proof.** Analogous to that of Proposition 2. \qed

Letting $B$ be $L^2(\pi)$ or $L_V$ yields some well-known result concerning $\chi^2$ or total variation distance (see, e.g., Kontoyiannis and Meyn, 2012). In particular, we have the following corollaries.

**Corollary 6.** Suppose that $B = L^2(\pi)$. Let $\mu$ be a probability measure dominated by $\Pi$, and suppose that $d\mu/d\Pi \in L^2(\pi)$. Denote by $\mu^P_k$ the distribution of $X_k$ when $X_0 \sim \mu$. Then for any $\rho > \rho(P - P_0)$, there exists $k_0 \in \mathbb{Z}_+$ such that whenever $k \geq k_0$,

$$\| \mu^P_k - \Pi \|_2 \leq \| \mu \|_2 \rho^k.$$ 

**Proof.** By Proposition 5, there exists $k_0$ such that whenever $k \geq k_0$,

$$\| \mu^P_k - \Pi \|_2 = \sup_{f \in L^2(\pi), \| f \|_2 = 1} \langle \frac{d\mu}{d\Pi}, (P^k - P_0)f \rangle \leq \| \frac{d\mu}{d\Pi} \|_2 \sup_{f \in L^2(\pi), \| f \|_2 = 1} \| (P^k - P_0)f \|_2 \leq \| \frac{d\mu}{d\Pi} \|_2 \rho^k.$$ 

The result follows by noting that

$$\| \frac{d\mu}{d\Pi} \|_2 = \sup_{f \in L^2(\pi), \| f \|_2 = 1} \langle \frac{d\mu}{d\Pi}, f \rangle = \| \mu \|_2.$$ 

\qed

**Remark 7.** If the chain is reversible, i.e., $\Pi(dx)p(x, dx') = \Pi(dx')P(x', dx)$, then the above result can be strengthened. To be precise, $k_0$ can be taken as 1, and $\rho$ can be taken as $\rho(P - P_0) = \| P - P_0 \|_2$ (Roberts and Rosenthal, 1997).

**Corollary 8.** Suppose that $B = L_V$ for some $V \geq 1$. Then for any $\rho > \rho(P - P_0)$, there exists $k_0 \in \mathbb{Z}_+$ such that whenever $k \geq k_0$,

$$\| P^k(x, \cdot) - \Pi(\cdot) \|_V \leq V(x)\rho^k$$

for (almost) all $x$.

**Proof.** Note that

$$\sup_{x \in \mathcal{X}} \frac{\| P^k(x, \cdot) - \Pi(\cdot) \|_V}{V(x)} = \sup_{x \in \mathcal{X}} \sup_{f \in L_V, \| f \|_V = 1} \int_{\mathcal{X}} f(x')(V(x')^{-1} \int_{\mathcal{X}} f(x')(P^k(x, dx') - P_0(x, dx'))) \ d\mu(x').$$
Exchanging the supremums on the right hand side yields
\[ \sup_{x \in X} \frac{\|P^k(x, \cdot) - \Pi(\cdot)\|_V}{V(x)} \leq \sup_{f \in L_V, \|f\|_V = 1} \|P^k - P_0\|_V f, \]
which by Proposition 5, is dominated by \( \rho^k \) whenever \( k \geq k_0 \) for some \( k_0 \).

All results above imply that one can bound the convergence rate of a chain if he/she can somehow find \( P \)'s spectrum (or at least \( \rho(P - P_0) \)). This, in general, is very difficult. In the next section we briefly review developments in spectral analysis of Markov operators, and take a look at some of the open problems in the field.

4 Spectral analysis of Markov operators

4.1 Geometric bounds by a path argument

The path argument (see, e.g., Diaconis and Stroock, 1991) is a method used mostly by computer scientists to bound \( \rho(P - P_0) \) for Markov chains on finite graphs. The bounds are called “geometric” because they depend on the geometric features of the graphs.

A graph \((V, E)\) is a set of vertices connected by edges. For simplicity assume that each pair of vertices are either not connected, or connected by a single edge. The set of vertices \( V \) forms a state space \( X \), on which a Markov chain can be defined in the following way. For two connected vertices, \( i \) and \( j \), \( p(i, j) \) is the probability that \( X_1 = j \) given that \( X_0 = i \). The \( p(i, j) \)'s form an Mtm, whose spectrum can be calculated directly. However, computer scientists are interested in the asymptotic behavior of \( \rho(P - P_0) \) as the size of the graph grows. Therefore, they look for analytic bounds on \( \rho(P - P_0) \) that are easy to analyze.

Assume that the chain is reversible, i.e., \( p(i, j)\pi(i) = p(j, i)\pi(j) \) for all \( (i, j) \in X^2 \). To construct the bound, one first chooses for each pair of vertices, \( (i, j) \), a path that connects the two, \( \gamma_{ij} \), which is a sequence of connecting edges. The length of each edge is defined to be the reciprocal of \( p(i, j)\pi(i) \pi(j) \), where \( (i, j) \) are its two endpoints. The length of a path is the sum of the lengths of all its edges. For each edge \( e \), consider all the paths that passes through it, that is, all \( (i, j) \)'s such that \( e \in \gamma_{ij} \). Then \( \rho(P - P_0) \) is bounded above by \( 1 - 1/\kappa \), where
\[ \kappa = \max_{e \in E} \sum_{(i,j) : e \in \gamma_{ij}} |\gamma_{ij}| \pi(i)\pi(j), \]
and \( |\gamma_{ij}| \) is the length of the path connecting \( (i, j) \).

If the paths are well-chosen, the above bound may be sharp enough to reflect the asymptotic behavior of \( \rho(P - P_0) \) as the graph grows in size. This feature can be intriguing in statistics, especially with the booming
of “big data”. However, it’s not clear if bounds of this type can be generalized to suit continuous state spaces.

4.2 Cheeger’s inequality

For a measurable set $A \subset \mathcal{X}$, the quantity

$$\int_A P(x, A^c)\Pi(dx)$$

is the probability of the chain moving from $A$ to its complement in one step assuming that it’s stationary. Roughly speaking, a rapidly-mixing chain should be capable of moving from a set to its complement with a relatively high probability. When $0 < \Pi(A) < 1$, let

$$k(A) = \frac{\int_A P(x, A^c)\Pi(dx)}{\Pi(A)\Pi(A^c)},$$

and define

$$k = \inf_{0<\Pi(A)<1} k(A).$$

Cheeger’s inequality states for that reversible chains,

$$1 - k \leq \rho(P - P_0) \leq 1 - k^2/8$$

(see, e.g., Lawler and Sokal, 1988).

One can verify that

$$1 - k = \sup_{f \in \mathcal{A}, f \neq 0} \frac{\langle f, (P - P_0)f \rangle}{\|f\|_2^2},$$

where $\mathcal{A}$ consists functions of the form $f(x) = 1_A(x) - \Pi(A)$. For reversible chains, $P - P_0$ is self-adjoint, thus,

$$\rho(P - P_0) = \|P - P_0\|_2 = \sup_{f \in L^2(\pi), f \neq 0} \frac{\langle f, (P - P_0)f \rangle}{\|f\|_2^2}.$$ 

This quantity is usually impossible to calculate due to the enormous size of $L^2(\pi)$. But Cheeger’s inequality tells us that one can often restrict his/her attention to $\mathcal{A}$, which consists of only shifted identity functions. In some simple examples, $k$ is actually possible to calculate. (See Belloni and Chernozhukov (2009) for a rather impressive application.) There aren’t many examples that an upper bound on $\rho(P - P_0)$ can be constructed in this way for practically relevant Markov chains. It’d be interesting to see if that will change in the near future.

Although $k$ can be difficult to calculate, an upper bound on $k$ is sometimes available, e.g., $k(A)$ for some $A \subset \mathcal{X}$. Such upper bounds yield lower bounds on $\rho(P - P_0)$, which are sometimes useful, especially when one wishes to show that a certain Markov chain converges slowly (see, e.g., Johndrow et al., 2016).
4.3 Compact operators and matrix approximation

It’s usually not possible to find the spectrum for a Markov chain defined on a continuous state space. However, there is some hope for estimating it when the operator is compact. A compact operator has a pure eigenvalue spectrum. Moreover, on Hilbert spaces, a linear operator is compact if and only if it can be approximated by finite-rank operators, e.g., matrices (see, e.g., Conway, 1990, Section 2.4). This is potentially very useful because the spectra of matrices are much easier to find than those of general linear operators. There have been some interesting works on spectral estimation for Markov operators along this line of reasoning (see, e.g., Koltchinskii and Giné, 2000; Adamczak and Bednorz, 2015; Qin et al., 2017).

Compactness is not always easy to establish. For Markov operators, the most common way to prove compactness is by showing they are Hilbert-Schmidt, or more often, trace-class, i.e., compact with absolutely summable eigenvalues. For reversible chains, the corresponding Markov operator is trace-class if and if \( \int_X P(x, dx) < \infty \), assuming that \( P(\cdot, \cdot) \) satisfies certain smoothness conditions. In recent years, many Markov operators corresponding to practically relevant MCMC algorithms have been shown to be trace-class via this formula (see, e.g., Choi and Hobert, 2013; Chakraborty and Khare, 2017; Pal et al., 2017).

Currently, most spectral approximation techniques are not quite practical because 1. the errors of the approximations can be very large, especially for chains on high dimensional state spaces, and 2. most of these methods require the operator to be Hilbert-Schmidt, which often does not hold. Suffice to say that there is still much work to be done in the area.

References


