Reversible Markov chains

Variational representations and ordering

Chris Sherlock

Abstract

This pedagogical document explains three variational representations that are useful when comparing the efficiencies of reversible Markov chains: (i) the Dirichlet form and the associated variational representations of the spectral gaps; (ii) a variational representation of the asymptotic variance of an ergodic average; and (iii) the conductance, and the equivalence of a non-zero conductance to a non-zero right spectral gap.

Introduction

This document relates to variational representations of aspects of a reversible Markov kernel, P, which has a limiting (hence, stationary) distribution, π. It is a record of my own learning, but I hope it might be useful to others. The central idea is the Dirichlet form, which can be thought of as a generalisation of expected squared jumping distance to cover all functions that are square-integrable with respect to π. The minimum (over all such functions with an expectation of 0 and a variance of 1) of the Dirichlet forms is the right spectral gap.

A key quantity of interest to researchers in MCMC is the asymptotic variance which relates to the the variance of an ergodic average, $\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right]$ as $n \to \infty$, and where $X_1, \ldots, X_n$ arise from successive applications of P, and f is some square-integrable function. A second variational representation, of $\langle f, (I-P)^{-1} f\rangle$, allows us to relate a limit of the variance of an ergodic average to the Dirichlet form. Finally, a third variational quantity, the conductance of a Markov kernel is introduced, and is also related to the Dirichlet form and hence to the variance of an ergodic average.

I found these variational representations particular useful in two pieces of research that I performed, mostly in 2016. Sherlock et al. (2017) uses the first two representations, while Sherlock and Lee (2017) uses conductance. In both pieces of work the variational representations allowed us to compare pairs of Markov kernels; if one Markov kernel has a particular property, such as the variance of an ergodic average being a particular value, and if we can relate aspects of this Markov kernel, such as its Dirichlet form or its conductance, to a sec-

ond Markov kernel, then we can often obtain a bound on the property of interest for the second kernel. This is not the only use of variational representations; e.g. in Lawler and Sokal (1988) conductance is used directly to obtain bounds on the spectral gap of several discrete-statespace Markov chains.

The most natural framework for representing the kernel is that of a bounded, self-adjoint operator on a Hilbert space. I was almost entirely unfamiliar with this before I embarked on the above pieces of research and so I will start by setting down the key aspects of this.

Preliminaries

Hilbert Space

Let $L^2(\pi)$ be the Hilbert space of real functions that are square integrable with respect to some probability measure, π:

$$\int \pi(\mathrm{d}x) f(x)^2 < \infty \Leftrightarrow f \in L^2(\pi),$$

equipped with the inner product (finite through Cauchy-Schwarz) and associated norm:

$$\langle f, g \rangle = \int \pi(\mathrm{d}x) f(x) g(x), \quad ||f||^2 = \langle f, f \rangle .$$

Let $L_0^2(\pi) \subset L^2(\pi)$ be the Hilbert space that uses the same inner product but includes only functions with $\mathbb{E}_\pi [f] = 0$:

$$\int \pi(\mathrm{d}x) f(x) = \langle f, 1 \rangle = 0.$$

For these functions, $\langle f, g \rangle = \mathrm{Cov} [f, g]$ and $\langle f, f \rangle = \mathrm{Var}_\pi [f]$.

Markov kernel and detailed balance

Let $\{X_t\}_{t=0}^{\infty}$ be a Markov chain on a statespace X with a kernel of $P(x, \mathrm{d}y)$ which satisfies detailed balance with respect to π:

$$\pi(\mathrm{d}x) P(x, \mathrm{d}y) = \pi(\mathrm{d}y) P(y, \mathrm{d}x).$$

For any measure, ν, we define

$$\nu P := \int_x \nu(\mathrm{d}x) P(x, \mathrm{d}y)$$

Then

$$\pi P = \int_x \pi(\mathrm{d}x) P(x, \mathrm{d}y) = \int_x \pi(\mathrm{d}y) P(y, \mathrm{d}x) = \pi,$$

so π is stationary for P.

P is bounded and self adjoint

Given a kernel (or 'operator') P we use the shorthand:

$$Pf(x) = (Pf)(x) := \int P(x, \mathrm{d}y) f(y).$$

Jensen's inequality gives

$$[(Pf)(x)]^2 = \left[\int P(x, \mathrm{d}y) f(y)\right]^2 \leq \int P(x, \mathrm{d}y) f(y)^2 = (Pf^2)(x).$$

Hence, because π is stationary for P,

$$||Pf||^2 = \int \pi(\mathrm{d}x)[(Pf)(x)]^2 \leq \int \pi(\mathrm{d}x)(Pf^2)(x) = \iint \pi(\mathrm{d}x) P(x, \mathrm{d}y) f^2(y) = \int \pi(\mathrm{d}y) f^2(y) = ||f||^2.$$

Thus $||Pf||/||f|| \leq 1$, and P is a bounded linear operator.

Further, if P satisfies detailed balance with respect to π

$$\begin{aligned} \langle Pf, g\rangle &= \iint \pi(\mathrm{d}x) P(x, \mathrm{d}y) f(y) g(x) = \iint \pi(\mathrm{d}y) P(y, \mathrm{d}x) f(y) g(x) \\ &= \iint \pi(\mathrm{d}x) P(x, \mathrm{d}y) f(x) g(y) = \langle f, Pg\rangle\,; \end{aligned}$$

P is self-adjoint.

The spectrum of a bounded, self–adjoint operator

The spectrum of P in $\mathcal{H}$ is $\{\rho : P - \rho I$ is not invertible in $\mathcal{H}\}$[1]; the spectrum of a bounded, self-adjoint operator is (accept it, but see below) a closed, bounded set on the real line; let

[1] i.e. there is at least one $g \in \mathcal{H}$ such that there is no *unique* $f \in \mathcal{H}$ with $(P - \rho I)f = g$

the upper and lower bounds be λ_{max} and λ_{min}. When P is a self-adjoint Markov kernel and $\mathcal{H} = L^2(\pi)$, $\lambda_{max} = 1$ and $\lambda_{min} \geq -1$. The spectral decomposition theorem for bounded, self-adjoint operators states that there is a finite positive measure, $a^*(\mathrm{d}\lambda)$ with support contained in the real interval $[\lambda_{min}, \lambda_{max}]$ such that

$$\langle f, P^n f \rangle = \int_{\lambda_{min}}^{\lambda_{max}} \lambda^n a^*(\mathrm{d}\lambda).$$

This decomposition is used twice hereafter; however it may be unfamiliar so the remainder of this section provides an intuition in terms of the eigenfunctions and eigenvalues of P.

Let P be a bounded, self-adjoint operator. A right eigenfunction e of P is a function that satisfies $Pe = \lambda e$ for some scalar, λ, the corresponding eigenvalue. Let $e_0(x), e_1(x), \ldots$ be a set of eigenfunctions of P, scaled so that $||e_i||^2 = \langle e_i, e_i \rangle = 1$, and with corresponding eigenvectors $\lambda_0, \lambda_1, \ldots$. Since, by definition, $(P - \lambda_i I)e_i = 0$, the spectrum is a superset of the set of eigenvalues. The intuition below comes from the case where the spectrum is precisely the set of eigenvalues, and the eigenfunctions span $\mathcal{H}$.

1. Just as for the eigenvectors of a finite self-adjoint matrix, it is possible to choose the eigenfunctions such that $\langle e_i, e_j \rangle = 0$ $(i \neq j)$;
2. moreover, as with self-adjoint matrices, all of the eigenvalues are real.
3. Furthermore, since P is bounded, all of the eigenvalues satisfy $|\lambda| \leq 1$.

If the eigenfunctions of an operator, P, span the Hilbert space, $\mathcal{H}$, for any $f \in \mathcal{H}$,

$$f = \sum_{i=0}^{\infty} a_i e_i,$$

where $a_i = \langle f, e_i \rangle$, and from which

$$\langle f, P^n f \rangle = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_i a_j \langle e_i, P^n e_j \rangle = \sum_{i=0}^{\infty} a_i^2 \lambda_i^n.$$

With $n = 0$ we obtain $\sum_{i=0}^{\infty} a_i^2 = ||f||^2 < \infty$. Thus, if the eigenfunctions span the space then a^* is discrete with mass a_i^2 at λ_i.

For a Markov kernel, $P1 = 1$, where 1 is the constant function and is, thus, a right eigenfunction with an eigenvalue of 1.

Spectral gaps and the Dirichlet form

Spectrum of P in $L_0^2(\pi)$; spectral gaps and geometric ergodicity

For any kernel P, we define $P^k f = P^{k-1} P f$, recursively.

Any function $f(x) \in L^2(\pi)$ can be written as $a_0 1 + f_0(x)$ where $f_0 \in L_0^2(\pi)$. Here $a_0 = \langle f, 1 \rangle = \mathbb{E}_\pi [f(X)]$. Thus

$$P^n f = \mathbb{E}_\pi [f(X)] + P^n f_0(x) \to \mathbb{E}_\pi [f(X)]$$

if P is ergodic. To bound the size of the remainder term, consider the spectrum of P restricted to functions in $L_0^2(\pi)$. This must be confined to $[\lambda_0^{min}, \lambda_0^{max}]$ with $-1 \leq \lambda_0^{min} \leq \lambda_0^{max} \leq 1$, where (by definition)

$$\lambda_0^{max} := \sup_{f \in L_0^2(\pi)} \frac{\langle f, Pf \rangle}{\langle f, f \rangle} = \sup_{f \in L_0^2(\pi): \langle f, f \rangle = 1} \langle f, Pf \rangle \quad \text{and} \quad \lambda_0^{min} := \inf_{f \in L_0^2(\pi): \langle f, f \rangle = 1} \langle f, Pf \rangle .$$

Let $\overline{\lambda} = \max(|\lambda_0^{min}|, |\lambda_0^{max}|)$. The (squared) size of the remainder is then

$$||P^n f_0||^2 = \langle f_0 P^n, P^n f_0 \rangle = \langle f_0, P^{2n} f_0 \rangle = \int_{\lambda_0^{min}}^{\lambda_0^{max}} \lambda^{2n} a^*(\mathrm{d}\lambda) \leq \overline{\lambda}^{2n} \int_{\lambda_0^{min}}^{\lambda_0^{max}} a^*(\mathrm{d}\lambda) = \overline{\lambda}^{2n} ||f||^2.$$

Thus $||P^n f|| \to 0$ geometrically quickly provided $\overline{\lambda} < 1$; i.e., provided the inequality is strict. P is then called geometrically ergodic. The right spectral gap is $\rho^{right} := 1 - \lambda_0^{max}$ and the left spectral gap is $\rho^{left} := 1 + \lambda_0^{min}$. Both must be non-zero for geometric ergodicity. Henceforth, for notational simplicity, we drop the subscript in $\lambda_0^{max/min}$.

Aside: when one or both of the spectral gaps is zero (e.g. the spectrum is $[-1, 1]$), then for any fixed n we can always find functions, f, with $||f|| = 1$, (albeit 'fewer and fewer' as $n \to \infty$) such that $||P^n f|| > 0.1$, say.

Dirichlet form, $\mathcal{E}_P(f)$

The concept of a spectral gap motivates the Dirichlet form for P and f,

$$\mathcal{E}_P(f) := \langle f, (I - P) f \rangle = \langle f, f \rangle - \langle f, Pf \rangle ,$$

since

$$\rho^{right} = \inf_{f \in L_0^2(\pi): \langle f, f \rangle = 1} \mathcal{E}_P(f) \quad \text{and} \quad \rho^{left} = 2 - \sup_{f \in L_0^2(\pi): \langle f, f \rangle = 1} \mathcal{E}_P(f).$$

Directly from the definition we have for two kernels P_1 and P_2 and some $\gamma > 0$:

$$\mathcal{E}_{P_1}(f) \geq \gamma \mathcal{E}_{P_2}(f) \; \forall \; f \in L_0^2(\pi) \Rightarrow \rho_{P_1}^{right} \geq \gamma \rho_{P_2}^{right}. \tag{1}$$

An alternative expression for the Dirichlet form provides a very natural intuition:

$$\begin{aligned} \mathcal{E}_P(f) &= \langle f, f \rangle - \langle f, Pf \rangle \\ &= \iint \pi(\mathrm{d}x) P(x, \mathrm{d}y) f(x)[f(x) - f(y)] \\ &= \iint \pi(\mathrm{d}x) P(x, \mathrm{d}y) f(y)[f(y) - f(x)] \\ &= \frac{1}{2} \iint \pi(\mathrm{d}x) P(x, \mathrm{d}y)[f(y) - f(x)]^2, \end{aligned} \tag{2}$$

where the penultimate line follows because P satisfies detailed balance with respect to π and the final line arises from the average of the two preceding lines. The Dirichlet form can, therefore, be thought of as a generalisation of expected squared jumping distance of the ith component of x, $\int \pi(\mathrm{d}x) P(x, \mathrm{d}y)(y_i - x_i)^2$, to consider the expected squared changes for any $f \in L^2(\pi)$.

Variance of an ergodic average

Suppose that we are interested in $\mathbb{E}_\pi[h(X)]$ for some $h \in L^2(\pi)$ and we estimate it by an average of the values in the Markov chain: $\hat{h}_n := \frac{1}{n} \sum_{i=1}^n P^{i-1} h(x)$. Typically $\mathrm{Var}\left[\hat{h}_n\right] \downarrow 0$ as $n \to \infty$, but scaling by $\sqrt{n}$ should keep it $\mathcal{O}(1)$.

We are, therefore, interested in

$$\mathrm{Var}(P, h) := \lim_{n \to \infty} \mathrm{Var}\left[\sqrt{n} \hat{h}_n\right].$$

So as to just consider mixing, we assume $X_0 \sim \pi$.

Without loss of generality we may assume $h \in L_0^2(\pi)$ (else just subtract its expectation). Since P is time-homogeneous,

$$\begin{aligned} \mathrm{Var}\left[\sum_{i=1}^n h(X_i)\right] &= \sum_{i=1}^n \mathbb{E}\left[h^2(X_i)\right] + 2 \sum_{i=1}^{n-1} \mathbb{E}\left[h(X_i), h(X_{i+1})\right] + 2 \sum_{i=1}^{n-2} \mathbb{E}\left[h(X_i), h(X_{i+2})\right] \\ &\quad + \cdots + 2\mathbb{E}\left[h(X_1), h(X_n)\right] \\ &= n\mathbb{E}\left[h^2(X)\right] + 2(n-1)\mathbb{E}\left[h(X_1), h(X_2)\right] + \cdots + 2\mathbb{E}\left[h(X_1), h(X_n)\right] \\ &= n \langle h, h \rangle + 2(n-1) \langle h, Ph \rangle + 2(n-2) \langle h, P^2 h \rangle + \cdots + 2 \langle h, P^{n-1} h \rangle. \end{aligned}$$

Using $(I-P)^{-1}$ to denote $I+P+P^2+P^3+\ldots$, we obtain [2]

$$\begin{aligned}\mathrm{Var}(P,h) &= \langle h,h\rangle + 2\sum_{i=1}^{\infty}\langle h,P^i h\rangle \\ &= 2\sum_{i=0}^{\infty}\langle h,P^i h\rangle - \langle h,h\rangle \\ &= 2\langle h,(I-P)^{-1}h\rangle - \langle h,h\rangle ,\end{aligned}$$

Writing $\langle h,h\rangle = \langle h,(I-P)(I-P)^{-1}h\rangle$ gives an equivalent form:

$$\mathrm{Var}(P,h) = \langle h,(I+P)(I-P)^{-1}h\rangle .$$

The spectral decomposition of P in $L_0^2(\pi)$ gives

$$\mathrm{Var}(P,h) = \int_{\lambda^{min}}^{\lambda^{max}} \frac{1+\lambda}{1-\lambda}\, a^*(\mathrm{d}\lambda) \leq \frac{1+\lambda^{max}}{1-\lambda^{max}} \int_{\lambda^{min}}^{\lambda^{max}} a^*(\mathrm{d}\lambda) = \frac{1+\lambda^{max}}{1-\lambda^{max}}\mathrm{Var}_\pi[h], \qquad (3)$$

where the inequality follows since $(1+\lambda)/(1-\lambda)$ is an increasing function of λ. The supremum of the spectrum of $L_0^2(\pi)$ provides an efficiency bound over all $h \in L_0^2(\pi)$.

Variance bounding kernels

The bound on the variance in (3) is in terms of $\lambda^{\max}$, not $\lambda^{\min}$. Even if there exists an eigenvalue $\lambda^{\min} = -1$, so the chain never converges (as opposed to the more usual case where the spectrum of P is $[-1,a]\cup 1$, but there is no eigenvalue at -1) the asymptotic variance can be finite. As an extreme example, consider the Markov chain with a transition matrix of

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

After $2n$ iterations this has been in state 1 exactly $1/2 = \pi(\{1\})$ of the time.

Roberts and Rosenthal (2008) realised that in almost all applications of MCMC it was the variance of the ergodic average that was important, and not (directly) convergence to the target. Specifically, it did not matter if $\rho^{left} = 0$. A kernel where $\rho^{right} > 0$ was termed variance bounding and this was shown to be equivalent to $\mathrm{Var}(P,f) < \infty$ for all $f \in L^2(\pi)$; i.e. if [hypothetical!] simple Monte Carlo would lead to the standard $\sqrt{n}$ rate of convergence of the ergodic average, a $\sqrt{n}$ rate would also be observed from the MCMC kernel.

[2] This is easy to see if the left and right spectral gaps of P are non-zero. When at least one gap is zero, for any $\beta < 1$, we have $\mathrm{Var}(\beta P,h) = \langle h,h\rangle + 2(1-1/n)\beta\langle h,Ph\rangle + 2(1-2/n)\beta^2\langle h,P^2h\rangle + \cdots + 2(1-(n-1)/n)\beta^{n-1}\langle h,P^{n-1}h\rangle \to \langle h,h\rangle + 2\sum_{i=1}^{\infty}\beta^i\langle h,P^ih\rangle < \infty$. Letting $\beta \uparrow 1$ then gives the result, even though $\mathrm{Var}(P,h)$ may be infinite.

Variational representation of $\langle f, Q^{-1} f \rangle$, and a key ordering

Firstly, notice that for real numbers f, g and q, $\sup_g 2fg - g^2 q$ is f^2/q, which is achieved when $g = f/q$. The same holds for the operator $Q = I - P$ when f and g are functions (see Appendix A for a proof):

$$\langle f, (I-P)^{-1} f \rangle = \sup_{g \in L_0^2(\pi)} 2\langle f, g \rangle - \langle g, (I-P)g \rangle = \sup_{g \in L_0^2(\pi)} 2\langle f, g \rangle - \mathcal{E}_P(g).$$

To see this, think of a reversible matrix, P and diagonalise it; the result for such matrices is just equivalent to the scalar result applied to each eigenvalue.

This leads to the alternative representation:

$$\mathrm{Var}(P, h) = \sup_{g \in L_0^2(\pi)} 4\langle h, g \rangle - 2\mathcal{E}_P(g) - \langle h, h \rangle .$$

Clearly, $\mathcal{E}_{P_1}(g) \geq \mathcal{E}_{P_2}(g) \; \forall \; g \in L_0^2(\pi) \Rightarrow \mathrm{Var}(P_1, h) \leq \mathrm{Var}(P_2, h)$, giving a *much* simpler proof of a result which was originally proved for finite-statespace Markov chains in Peskun (1973) and then generalised to general statespaces by Tierney (1998). But we can go further: suppose that $\mathcal{E}_{P_1}(g) \geq \gamma \mathcal{E}_{P_2}(g)$ for all $g \in L_0^2(\pi)$ and some $\gamma > 0$, then, for any $g \in L_0^2(\pi)$

$$2\langle h, g \rangle - \mathcal{E}_{P_1}(g) \leq 2\langle h, g \rangle - \gamma \mathcal{E}_{P_2}(g) = \frac{1}{\gamma} \left\{ 2\langle h, g_* \rangle - \mathcal{E}_{P_2}(g_*) \right\},$$

where $g_* := \gamma g \in L_0^2(\pi)$. So

$$\sup_{g \in L_0^2(\pi)} 2\langle h, g \rangle - \mathcal{E}_{P_1}(g) \leq \frac{1}{\gamma} \sup_{g \in L_0^2(\pi)} 2\langle h, g \rangle - \mathcal{E}_{P_2}(g),$$

and hence

$$\mathcal{E}_{P_1}(f) \geq \gamma \mathcal{E}_{P_2}(f) \; \forall \; f \in L_0^2(\pi) \Rightarrow \mathrm{Var}(P_1, h) + \langle h, h \rangle \leq \frac{1}{\gamma} \left\{ \mathrm{Var}(P_2, h) + \langle h, h \rangle \right\}. \tag{4}$$

This result appears as Lemma 32 in Andrieu et al. (2016) (see also Caracciolo et al., 1990).

Propose-accept-reject kernels

Many kernels consist of making a proposal, which is then either accepted or rejected:

$$P(x, \mathrm{d}y) := q(x, \mathrm{d}y)\alpha(x, y) + (1 - \overline{\alpha}(x))\delta(y - x),$$

where

$$\overline{\alpha}(x) := \int q(x, \mathrm{d}y)\alpha(x, y)$$

is the average acceptance probability from x. Then, from (2),

$$\mathcal{E}_P(f) = \frac{1}{2} \iint \pi(\mathrm{d}x)q(x, \mathrm{d}y)\alpha(x, y)\{f(y) - f(x)\}^2.$$

The right spectral gap is, therefore,

$$\rho^{right} = \frac{1}{2} \inf_{f \in L_0^2(\pi):\langle f,f\rangle=1} \iint \pi(\mathrm{d}x)q(x, \mathrm{d}y)\alpha(x, y)\{f(y) - f(x)\}^2.$$

There is a similar formula for the left spectral gap.

Results (1) and (4) then lead directly to the following: if two propose-accept-reject kernels, both reversible with respect to π, satisfy

$$q_1(x, \mathrm{d}y)\alpha_1(x, y) \geq \gamma q_2(x, \mathrm{d}y)\alpha_2(x, y) \quad \forall\ x, y$$

then $\rho_{P_1}^{right} \geq \gamma\rho_{P_2}^{right}$ and $\mathrm{Var}(P_1, h) + \mathrm{Var}_\pi[h] \leq \frac{1}{\gamma}\{\mathrm{Var}(P_2, h) + \mathrm{Var}_\pi[h]\}$.

From (4), if for some $\gamma > 0$ $\mathcal{E}_{P_1}(f) \geq \gamma\mathcal{E}_{P_2}(f)\ \forall\ f \in L_0^2(\pi)$ and P_2 is variance bounding, then so is P_1. Unfortunately it is rare that we can be sure of the ordering of the Dirichlet forms for all f; indeed we might be sure that there is no fixed γ for which the ordering does hold. When we cannot simply resort to a uniform ordering of Dirichlet forms then the elegant concept of conductance can come to our aid.

Conductance

Consider any measurable set, $\mathcal{A} \subseteq \mathsf{X}$. The conductance of $\mathcal{A}$ is defined as

$$\kappa_P(\mathcal{A}) := \frac{1}{\pi(\mathcal{A})} \iint_{x\in\mathcal{A},y\in\mathcal{A}^c} \pi(\mathrm{d}x)P(x, \mathrm{d}y),$$

where $\pi(\mathcal{A}) := \int_{\mathcal{A}} \pi(\mathrm{d}x)$. Loosely speaking, $\kappa_P(\mathcal{A})$ is the probability of moving to $\mathcal{A}^c$ conditional on the chain currently following the stationary distribution truncated to $\mathcal{A}$: $\pi(\mathrm{d}x)1_{x\in\mathcal{A}}$. Our analysis of the properties of $\kappa_P(\mathcal{A})$ will be via the symmetric quantity

$$\kappa_P^*(\mathcal{A}) = \kappa_P^*(\mathcal{A}^c) := \frac{1}{\pi(\mathcal{A})\pi(\mathcal{A}^c)} \int_{x\in\mathcal{A},y\in\mathcal{A}^c} \pi(\mathrm{d}x)P(x, \mathrm{d}y) = \frac{\kappa_P(\mathcal{A})}{\pi(\mathcal{A}^c)}.$$

The first equality follows because the kernel is reversible with respect to π, and this also implies that $\pi(\mathcal{A})\kappa_P(\mathcal{A}) = \pi(\mathcal{A}^c)\kappa_P(\mathcal{A}^c)$. Since $\kappa_P(\mathcal{A}^c) \leq 1$, $\kappa_P(\mathcal{A}) \leq \pi(\mathcal{A}^c)/\pi(\mathcal{A})$, which can be arbitrarily small. To define the conductance of the kernel P we therefore only consider sets $\mathcal{A}$ with $\pi(\mathcal{A}) \leq 1/2$:

$$\kappa_P := \inf_{\mathcal{A}:\pi(\mathcal{A})\leq 1/2} \kappa(\mathcal{A}).$$

No such restriction is required for:

$$\kappa_P^* := \inf_{\mathcal{A}} \kappa_P^*(\mathcal{A}).$$

Further, since $\pi(\mathcal{A}) \leq 1/2 \Rightarrow 1/2 \leq \pi(\mathcal{A}^c) \leq 1$, we have $\kappa_P \leq \kappa_P^* \leq 2\kappa_P$.

Setting $f_{\mathcal{A}}(x) = [1_{x\in\mathcal{A}} - \pi(A)]/\sqrt{\pi(A)\pi(A^c)}$ (so that $f_{\mathcal{A}} \in L_0^2(\pi)$ and $\langle f_{\mathcal{A}}, f_{\mathcal{A}}\rangle = 1$) in the expression for the Dirichlet form (2) gives

$$\mathcal{E}_P(f_{\mathcal{A}}) = \frac{1}{2\pi(\mathcal{A})\pi(\mathcal{A}^c)}\left\{\int_{x\in\mathcal{A},y\in\mathcal{A}^c} \pi(\mathrm{d}x)P(x,\mathrm{d}y) + \int_{x\in\mathcal{A}^c,y\in\mathcal{A}} \pi(\mathrm{d}x)P(x,\mathrm{d}y)\right\} = \kappa_P^*(\mathcal{A}) = \frac{\kappa_P(\mathcal{A})}{\pi(\mathcal{A}^c)}.$$

So

$$\rho^{right} = \inf_{f\in L_0^2(\pi):\langle f,f\rangle=1} \mathcal{E}_P(f) \leq \inf_{\mathcal{A}} \mathcal{E}_P(f_{\mathcal{A}}) = \kappa_P^* \leq 2\kappa_P. \tag{5}$$

Hence, if P has a right spectral gap (so it is variance bounding) then its conductance is non-zero. Amazingly, the converse is also true: if the conductance of P is non-zero then P has a right spectral gap:

$$\rho^{right} \geq \frac{\kappa_P^{*2}}{2} \geq \frac{\kappa_P^2}{2}. \tag{6}$$

Thus, non-zero conductance is equivalent to a non-zero right spectral gap is equivalent to variance bounding. Equations (5) and (6) together are sometimes called Cheeger bounds.

A proof (6) for finite Markov chains is given in Diaconis and Stroock (1991). A more accessible and, as far as I can see, more general, proof of the looser inequality, that $\rho^{right} \geq \kappa_P^2/8$ is given in Lawler and Sokal (1988). This has a 'standard bit' which uses the Cauchy-Schwarz inequality to obtain an inequality for ρ^{right}, a 'beautiful bit' which relates the expression from the 'standard bit' to conductance, and then an 'ugly bit', which proves that the final expression is always greater than 0 if $\kappa_P > 0$. I will follow Lawler and Sokal (1988) for the first two parts and then provide a neater solution to the final part.

Denote the symmetric measure $\pi(\mathrm{d}x)P(x,\mathrm{d}y)$ by $\mu(\mathrm{d}x,\mathrm{d}y)$. We also set $g(x) = f(x) + c$ for some (currently) arbitrary real constant, c.

The standard bit. Then Cauchy-Schwarz (twice) and the symmetry of μ gives:

$$\begin{aligned}
\mathbb{E}_\mu \left[|g(X)^2 - g(Y)^2|\right]^2 &= \mathbb{E}_\mu \left[|g(X) - g(Y)|\ |g(X) + g(Y)|\right]^2 \\
&\leq \mathbb{E}_\mu \left[\{g(X) - g(Y)\}^2\right] \mathbb{E}_\mu \left[\{g(X) + g(Y)\}^2\right] \\
&\leq \mathbb{E}_\mu \left[\{f(X) - f(Y)\}^2\right] 2\mathbb{E}_\mu \left[g(X)^2 + g(Y)^2\right] \\
&= 4\mathbb{E}_\mu \left[\{f(X) - f(Y)\}^2\right] \mathbb{E}_\pi \left[g(X)^2\right].
\end{aligned}$$

So

$$\mathcal{E}_P(f) \geq \frac{1}{8\mathbb{E}_\pi \left[g(X)^2\right]} \mathbb{E}_\mu \left[|g(X)^2 - g(Y)^2|\right]^2.$$

The beautiful bit. We now relate the numerator of the above expression to the conductance. The proof is symmetrical, the first half manipulates the denominator so that conductance may be used, with the second half reversing the route of the first.

Set $\mathcal{A}_t := \{x : g(x)^2 \leq t\}$. Then by the symmetry of μ,

$$\begin{aligned}
\int_{\mathsf{X}\times\mathsf{X}} \mu(\mathrm{d}x, \mathrm{d}y)|g(y)^2 - g(x)^2| &= 2\int_{\mathsf{X}\times\mathsf{X}} \mu(\mathrm{d}x, \mathrm{d}y) 1_{g(x)^2 < g(y)^2} \{g(y)^2 - g(x)^2\} \\
&= 2\int_{\mathsf{X}\times\mathsf{X}} \mu(\mathrm{d}x, \mathrm{d}y) \int_{t=0}^{\infty} \mathrm{d}t 1_{g(x)^2 \leq t < g(y)^2} \\
&= 2\int_{t=0}^{\infty} \mathrm{d}t \int_{\mathsf{X}\times\mathsf{X}} \mu(\mathrm{d}x, \mathrm{d}y) 1_{g(x)^2 \leq t < g(y)^2} \\
&= 2\int_{t=0}^{\infty} \mathrm{d}t \int_{\mathcal{A}_t \times \mathcal{A}_T^c} \mu(\mathrm{d}x, \mathrm{d}y) \\
&\geq 2\kappa_P^* \int_{t=0}^{\infty} \mathrm{d}t \int_{\mathcal{A}_t \times \mathcal{A}_t^c} \pi(\mathrm{d}x)\pi(\mathrm{d}y) \\
&= \kappa_P^* \int_{\mathsf{X}\times\mathsf{X}} \pi(\mathrm{d}x)\pi(\mathrm{d}y)|g(y)^2 - g(x)^2|.
\end{aligned}$$

So

$$\mathcal{E}_P(f) \geq \frac{\kappa_P^{*2} \mathbb{E}_{\pi\times\pi} \left[|g(X)^2 - g(Y)^2|\right]^2}{8\mathbb{E}_\pi \left[g(X)^2\right]}.$$

But since this is true for all c, we have

$$\rho^{right} \geq \frac{\kappa_P^{*2}}{8} \inf_{f \in L_0^2(\pi): \langle f, f\rangle = 1} \sup_c \frac{\mathbb{E}_{\pi\times\pi} \left[|\{f(Y) + c\}^2 - \{f(X) + c\}^2|\right]^2}{\mathbb{E}_\pi \left[\{f(X) + c\}^2\right]}.$$

A less ugly bit. Since $\mathbb{E}_\pi \left[\{f(X) + c\}^2\right] = 1 + c^2$, we need to show that

$$\inf_{f \in L_0^2(\pi): \langle f, f\rangle = 1} \sup_c \frac{\mathbb{E}_{\pi\times\pi} \left[|\{f(Y) + c\}^2 - \{f(X) + c\}^2|\right]}{\sqrt{1 + c^2}} \geq 1.$$

Let $A = f(X)$ and $B = f(Y)$ be independent and identically distributed with an expectation of 0 and a variance of 1. We are interested in

$$\frac{\mathbb{E}\left[|\{A+c\}^2 - \{B+c\}^2|\right]}{\sqrt{1+c^2}}. \tag{7}$$

Below, I will show that

$$\mathbb{E}\left[|A^2 - B^2|\right] + 4\mathbb{E}\left[|A-B|\right]^2 \geq 2. \tag{8}$$

Then, as in Lawler and Sokal (1988), consider letting $c \to \infty$ or setting $c = 0$. With the former:

$$\lim_{c\to\infty} \frac{\mathbb{E}\left[|\{A+c\}^2 - \{B+c\}^2|\right]}{\sqrt{1+c^2}} = 2\mathbb{E}\left[|A-B|\right],$$

so if $\mathbb{E}\left[|A-B|\right] > 1/2$ we are done. If not, setting $c = 0$ in (7) and using (8) leaves us:

$$\mathbb{E}\left[|A^2 - B^2|\right] \geq 1.$$

To prove (8):

$$\begin{aligned}
\mathbb{E}\left[|A^2 - B^2|\right] &= \int_0^\infty \mathbb{P}\left(|A^2 - B^2| > t\right) \mathrm{d}t = \int_0^\infty \mathbb{P}\left(A^2 > t + B^2\right) + \mathbb{P}\left(B^2 > t + A^2\right) \mathrm{d}t \\
&= 2\int_0^\infty \mathbb{P}\left(B^2 > t + A^2\right) \mathrm{d}t = 2\mathbb{E}_A\left[\int_0^\infty \mathbb{P}\left(B^2 > t + A^2\right) \mathrm{d}t | A\right].
\end{aligned}$$

But

$$\begin{aligned}
\int_0^\infty \mathbb{P}\left(B^2 > t + a^2\right) \mathrm{d}t &= \int_{a^2}^\infty \mathbb{P}\left(B^2 > v\right) \mathrm{d}v = \int_0^\infty \mathbb{P}\left(B^2 > v\right) \mathrm{d}v - \int_0^{a^2} \mathbb{P}\left(B^2 > v\right) \mathrm{d}v \\
&= 1 - \int_0^{|a|} 2u\mathbb{P}\left(|B| > u\right) \mathrm{d}u \\
&\geq 1 - 2|a| \int_0^\infty \mathbb{P}\left(|B| > u\right) = 1 - 2|a|\mathbb{E}\left[|B|\right].
\end{aligned}$$

Combining the two end results gives $\mathbb{E}\left[|A^2 - B^2|\right] \geq 2 - 4\mathbb{E}\left[|A|\right]\mathbb{E}\left[|B|\right]$; i.e.

$$\mathbb{E}\left[|A^2 - B^2|\right] + 4\mathbb{E}\left[|A|\right]^2 \geq 2.$$

However Jensen's inequality provides: $\mathbb{E}\left[|A-B|\right] = \mathbb{E}\left[\mathbb{E}\left[|A-B|\right]|A\right] \geq \mathbb{E}\left[|A - \mathbb{E}\left[B\right]|\right] = \mathbb{E}\left[|A|\right]$, and (8) follows.

Acknowledgements

I am grateful to Dr. Daniel Elton for providing the proof in Appendix A and Dr. Dootika Vats for spotting an error in an earlier version of this document.

A Variational representation of $\langle f, Q^{-1} f \rangle$

Let $\mathcal{H}$ be a Hilbert space and let Q be a positive operator on $\mathcal{H}$ (i.e., an operator that has a square root). Then for $f \in \mathcal{H}$,

$$\langle f, Q^{-1} f \rangle = \sup_{g \in \mathcal{H}} 2\mathrm{Re}(\langle f, g \rangle) - \langle g, Qg \rangle \,.$$

(Our Hilbert space is real, so we do not need the Re() function.)

Proof Let $\phi = Q^{-1} f$, so $f = Q\phi$. The left-hand side is then

$$\begin{aligned}
\langle Q\phi, \phi \rangle &= \langle Q^{1/2}\phi, Q^{1/2}\phi \rangle = ||Q^{1/2}\phi||^2 \\
&\geq ||Q^{1/2}\phi||^2 - ||Q^{1/2}(\phi - g)||^2 \\
&= 2\mathrm{Re}(\langle Q^{1/2}\phi, Q^{1/2} g \rangle) - ||Q^{1/2} g||^2 \\
&= 2\mathrm{Re}(\langle Q\phi, g \rangle) - \langle g, Qg \rangle \\
&= 2\mathrm{Re}(\langle f, g \rangle) - \langle g, Qg \rangle \,.
\end{aligned}$$

The construction of the inequality shows that the supremum is achieved at $g = Q^{-1} f$.

References

Andrieu, C., Lee, A., and Vihola, M. (2016). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*. to appear.

Caracciolo, S., Pelissetto, A., and Sokal, A. D. (1990). Nonlocal monte carlo algorithm for self-avoiding walks with fixed endpoints. *Journal of Statistical Physics*, 60(1):1–53.

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.*, 1(1):36–61.

Lawler, G. F. and Sokal, A. D. (1988). Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality. *Trans. Amer. Math. Soc.*, 309(2):557–580.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60:607–612.

Roberts, G. O. and Rosenthal, J. S. (2008). Variance bounding Markov chains. *Ann. Appl. Probab.*, 18(3):1201–1214.

Sherlock, C. and Lee, A. (2017). Variance bounding of delayed-acceptance kernels. *ArXiv e-prints*.

Sherlock, C., Thiery, A., and Lee, A. (2017). Pseudo-marginal Metropolis–Hastings using averages of unbiased estimators. *Biometrika*. Accepted subject to minor revisions.

Tierney, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *Ann. Appl. Probab.*, 8(1):1–9.