Markov chains with a large finite state space

Markov chains with a large finite state space

May 9, 2003

Contents

1 Introduction

It is a well known fact that, under suitable conditions, the distribution of a Markov chain on a finite state space converges to equilibrium as time increases. The main focus of this essay is on techniques that may be useful in capturing how this convergence is affected asymptotically as $N \to \infty$, where N is some kind of scaling parameter for the state space of the Markov chain.

There is some "real world" motivation for the topics that are covered in this essay and various applications are introduced briefly in section 6. In particular, there is discussion covering two implementations of the Metropolis algorithm and also how Markov chain techniques may be used in the generation of random numbers. The large state space setting is applicable to these examples since they both arise from a computational environment.

The fact that many of the applications are computational means that it is discrete time chains that are of greatest interest and it is for this reason that the majority of the examples given are for discrete chains. However, it should be noted that most of the results do have parallels in continuous time and section 3 touches on some of these.

We now introduce the basic notation that will be used in the essay. The state space shall be denoted by S and in general we have $|S| = N$. K represents an aperiodic, irreducible Markov operator or transition matrix throughout, with kernel $K(x, y)$ defined by

$$Kf(x) = \sum_y K(x,y)f(y).$$

It is a fact that, since K is irreducible, there exists a unique probability measure π such that $\pi K = \pi$. This formulation means that π is the invariant distribution for K. It should be noted that there are other ways of defining π leading to different names for it. In the essay, invariant distribution, equilibrium distribution and stationary distribution are used interchangeably. The aperiodicity allows us to infer that

$$K^n(x,y) \to \pi(y) \text{ as } n \to \infty, \forall x, y \in S. \tag{1}$$

That this limit occurs is common knowledge and it is the rate of convergence of $K^n(x,\cdot)$ to π that is our primary concern. Consequently, we need a way of measuring the distance between the two distributions. The following section introduces the distances that are used in the essay and explains briefly why they are of interest.

1.1 Choice of distance

A natural choice of distance between probability measures is the (total) variation distance and is defined for measures μ and π on S to be

$$||\mu - \pi||_{TV} = \frac{1}{2}\sum_{x\in S} |\mu(x) - \pi(x)|.$$

It is this distance that we shall be most interested in when considering how far $K^n(x, \cdot)$ is from π.

Other distances regularly used in analysis are the l^p distances. These are induced by the l^p norm with respect to the invariant measure π:

$$||f||_p = \left(\sum_{x\in S} |f(x)|^p \pi(x)\right)^{1/p}.$$

It is the l^2 distance that is of most use in the areas that are discussed in this essay. This is because of the Hilbert space structure that it provides, the relevant inner product being defined by

$$\langle f, g\rangle = \sum_x f(x)g(x)\pi(x).$$

When using this distance, rather than working with $K^n(x, \cdot)$, it becomes sensible to work with the density k_x^n, defined by

$$k_x^n(y) = \frac{K^n(x,y)}{\pi(y)}.$$

The reason for this is one of scaling as it allows a comparison with the variation distance. By definition, we have that

$$2||K^n(x,\cdot) - \pi||_{TV} = ||k_x^n - 1||_1$$

and we also have

$$||k_x^n - 1||_1 \leq ||k_x^n - 1||_2 \leq \pi_\star^{-1/2}||k_x^n - 1||_1$$

where $\pi_\star := \min_x \pi(x)$. As we shall see in the course of this essay the left hand inequality is particularly useful as it allows us to obtain upper bounds on the variation distance using l^2 arguments.

Another quantity of interest is the separation distance. This is defined to be, for measures μ and π,

$$d_{sep}(\mu, \pi) = \max_x \left\{1 - \frac{\mu(x)}{\pi(x)}\right\}.$$

We note that, although this is not a metric, it is useful to us because it satisfies

$$||\mu - \pi||_{TV} \leq d_{sep}(\mu, \pi)$$

and so any upper bound for d_{sep} immediately gives us a bound for the variation distance.

1.2 Methods for bounding distance from stationarity

There is a range of methods for bounding the distance from stationarity of Markov chains. These include using Fourier analysis, coupling times, strong stationary times and eigenvalue bounds. Fourier analysis is particularly applicable to chains on finite groups and we shall see an example of its use in section 6.2.

To contrast with the eigenvalue methods that will be the core of this essay, a description of the ideas behind coupling and strong stationary time bounds is given here. First, we consider two chains, one starting from $x \in S$ and another starting with initial distribution π. Define T to be the time that the two chains meet (the *coupling time*). It can easily be shown that, $\forall n \geq 1$,

$$||K^n(x, \cdot) - \pi||_{TV} \leq P(T > n). \tag{2}$$

In fact, in [11], it is shown how this type of coupling argument may be used to prove the convergence to equilibrium stated at (1). In section 6.2 an example of how an adapted version of this bound may be used is presented.

Define $(X_n)_0^\infty$ to be a Markov chain on S with transition matrix K, starting from $x \in S$. A *strong stationary time* is defined to be a stopping time for $(X_n)_0^\infty$, T say, such that

$$P(X_n = y | T = n) = \pi_y, \ \forall \ 0 \leq n < \infty, y \in S.$$

i.e.(conditional on T being finite) X_T has the distribution π. If such a time exists it can then be deduced that, $\forall n \geq 0$,

$$d_{sep}(K^n(x, \cdot), \pi) \leq P(T > n).$$

It is shown in [2] that strong stationary times are actually special cases of coupling times. However, the intuitive description is different and does give another way of approaching the construction of an appropriate time.

Furthermore, it has been proved that there exist optimal coupling and strong stationary times (ones for which the inequalities are equalities). However, one major problem with using such stopping times to find bounds is that there is no general theory about how to construct them.

It transpires that upper bounds on the distance from stationarity may be constructed using knowledge of the eigenvalues of K only. This approach is introduced in section 2. Although we do lose the optimality of the bounds, we are able to get away with using less information about the chain. This means that bounds of this nature are of particular interest when dealing with a large state space.

Since the eigenvalues of K are of such interest, there has been progress made in trying to estimate their values by means other than explicit calculation. One approach is introduced in section 4 and involves the consideration of geometric quantities associated with the Markov chain. Another approach that is not touched on here is to use comparison techniques. The basic idea for these is that we may sometimes deduce things about a chain from a similar chain that is easier to analyse by making suitable comparisons.

In section 5 we discuss an aspect of convergence to equilibrium that is not captured by looking at upper bounds alone. By analysing the convergence more closely we see that in some chains the distance from stationarity remains large for some time and then rapidly decays. This is known as the cutoff (or threshold) phenomenon.

2 Eigenvalue bounds

The time reversal of K is the Markov operator K^* defined by

$$K^*(x, y) = \frac{\pi(y)K(y, x)}{\pi(x)}.$$

Note that this means K^* is the adjoint of K with respect to $\langle \cdot, \cdot \rangle$. If $K^* = K$ we say that K is *reversible*. It turns out that reversibility is a property that gives us an immediate and simple way to obtain exponential bounds on the distance from stationarity of a Markov chain. In this section we shall see why the eigenvalues of K are of particular interest in this case. We shall also see how, in the irreversible case, one may establish similar types of bound by considering a reversible Markov operator related to K.

2.1 Characterisation of eigenvalues of reversible K

When K is reversible or, equivalently, self-adjoint then it is a result from linear mathematics that it has real eigenvalues $(\beta_i)_{i=0}^{N-1}$ and that there exists a corresponding basis of real orthonormal eigenfunctions $(\psi_i)_{i=0}^{N-1}$. For convenience, arrange the eigenvalues in non-increasing order. Note then that $\beta_0 = 1$ and $\psi_0 \equiv 1$.

There is more than one way to characterise the eigenvalues of K, the following minimax/maximin representations prove to be useful:

$$\begin{aligned} 1-\beta_i &= \min_{W:\dim(W)\geq i+1}\left\{\max_{f\in W, f\neq 0}\left\{\frac{\varepsilon_K(f,f)}{\|f\|_2^2}\right\}\right\} \\ &= \max_{W:\dim(W^{\perp})\leq i}\left\{\min_{f\in W}\left\{\frac{\varepsilon_K(f,f)}{\|f\|_2^2}\right\}\right\} \end{aligned}$$

where

$$\varepsilon_K(f,f) := \langle (I-K)f, f\rangle.$$

As we shall see, there is particular interest in β_1 and β_{N-1} for which the above result specialises to

$$\beta_1 = 1 - \min_f\left\{\frac{\varepsilon_K(f,f)}{\|f\|_2^2}\colon\ \pi(f)=0\right\} \tag{3}$$

$$\beta_{N-1} = 1 - \max_f\left\{\frac{\varepsilon_K(f,f)}{\|f\|_2^2}\right\}. \tag{4}$$

2.2 Eigenvalue bounds for reversible chains

Suppose that K is reversible. Using the notation of the previous section, the orthonormality of the eigenvectors allows us to express K^n in terms of its componentwise projections onto each basis vector.

$$\begin{aligned} K^n(x,y) &= K^n 1_y(x) = \sum_{i=0}^{N-1}\langle K^n 1_y, \psi_i\rangle\psi_i(x) = \sum_{i=0}^{N-1}\langle 1_y, K^n\psi_i\rangle\psi_i(x) \\ &= \sum_{i=0}^{N-1}\langle 1_y, \psi_i\rangle\beta_i^n\psi_i(x) = \sum_{i=0}^{N-1}\beta_i^n\psi_i(x)\psi_i(y)\pi(y) \end{aligned} \tag{5}$$

This decomposition then allows us to evaluate precisely the l^2 distance to stationarity of the chain (using the orthonormality of the eigenfunctions):

$$\|k_x^n - 1\|_2^2 = \|\sum_{i=1}^{N-1}\beta_i^n\psi_i(x)\psi_i\|_2^2 = \sum_{i=1}^{N-1}\beta_i^{2n}\psi_i^2(x).$$

As stated in section 1.1, this result may be used to provide an upper bound for the total variation distance. However, this bound may be unsatisfactory as it depends on knowing the eigenfunctions of the Markov operator which could be difficult or time consuming to calculate. We can, rather crudely, remove this dependence by bounding $\psi_i^2(x)$ by $1/\pi(x)$ to give the bound

$$||k_x^n - 1||_2^2 \leq \frac{1}{\pi(x)} \sum_{i=1}^{N-1} \beta_i^{2n}.$$

This result still depends on knowing all of the eigenvalues of the chain. In cases where N is large then this may prove to be problematic. As we will see later, there exist methods for bounding the largest (in modulus) of these $\beta_\star = \min\{\beta_1, |\beta_{N-1}|\}$ and so we may want to use the bound

$$||k_x^n - 1||_2^2 \leq \beta_\star^{2n} \sum_{i=1}^{N-1} \psi_i^2(x) = \beta_\star^{2n} ||k_x^0 - 1||_2^2. \tag{6}$$

We have

$$||k_x^0 - 1||_2^2 = \sum_y |\frac{1_x(y)}{\pi(y)} - 1|^2 \pi(y) = \frac{1-\pi(x)}{\pi(x)} \leq \frac{1}{\pi(x)} \tag{7}$$

and hence, (6) becomes

$$||k_x^n - 1||_2^2 \leq \frac{1}{\pi(x)} \beta_\star^{2n}. \tag{8}$$

Although eigenvalue bounds can be sharp, it should be noted that by bounding the variation distance using this result we are indeed losing accuracy. The example (from [6]) we now give is one case where using just $\beta_\star$ is far from optimal.

Random walk on a group

First, let G be a finite group and P be a distribution on G. Define K by

$$K(x,y) = P(yx^{-1}).$$

Now define $(X_n)_0^\infty$ to be a random walk on G with step distribution P, i.e. $X_n = Y_n * Y_{n-1} * ... * Y_1$ and $X_0 \equiv g_0$, where g_0 is the identity element of G and $(Y_n)_0^\infty$ are independent random variables on G, each with distribution P. Thus, $(X_n)_0^\infty$ is a Markov chain on G with transition matrix K and $X_0 \equiv g_0$.

We now consider the specific distribution

$$P = (1-\theta)\delta_{g_0} + \theta u$$

where $\theta \in (0,1]$, δ_{g_0} is the point mass at the identity and u is the uniform distribution on G. With this setup we have that $\pi = u$, $\beta_\star = 1-\theta$ and

$$K^n(g_0, \cdot) = (1-\theta)^n \delta_{g_0} + (1-(1-\theta)^n)u.$$

From which it is straightforward to obtain that

$$||K^n(g_0,\cdot) - \pi||_{TV} = \left(1 - \frac{1}{|G|}\right)(1-\theta)^n.$$

The upper bound, (8), gives that

$$||K^n(g_0,\cdot) - \pi||_{TV} \leq \frac{1}{2}||k_x^n - 1||_2 \leq \frac{1}{2\sqrt{\pi(x)}}\beta_\star^n = \frac{1}{2}|G|^{1/2}(1-\theta)^n$$

which is not a good bound for large $|G|$.

2.3 Eigenvalue bounds for irreversible chains

The arguments used in the previous section clearly depend on the reversibility of K to allow the real orthonormal basis to be used. We now discuss methods of *reversiblization* presented by Fill, [8], which allow similar results to be obtained by constructing reversible Markov operators from K.

Define the *multiplicative reversiblization*, $M(K)$, of K by

$$M(K) = KK^*$$

and the *additive reversiblization*, $A(K)$, of K by

$$A(K) = \frac{1}{2}(K + K^*). \tag{9}$$

It is straightforward to show that both $A(K)$ and $M(K)$ are both reversible Markov operators with invariant distribution π. If we let $(\beta_i(M))_{i=0}^{N-1}$ be the eigenvalues of M arranged in non-increasing order then we obtain:

Theorem

$$||k_x^n - 1||_2^2 \leq \frac{1}{\pi(x)}\beta_1(M)^n$$

To prove this we require the following lemma:

Lemma (Mihail's Identity)

For K and $M(K)$ defined as above we have

$$\mathrm{Var}_\pi(f) = \mathrm{Var}_\pi(K^* f) + \varepsilon_M(f,f).$$

Proof of Lemma: If we first note that $\pi(f) = \pi(K^* f)$, then it follows that

$$\begin{aligned}\varepsilon_M(f,f) &= \langle (I-KK^*)f,f\rangle = \langle f,f\rangle - \langle K^*f,K^*f\rangle \\ &= \mathrm{Var}_\pi(f) - \mathrm{Var}_\pi(K^*f)\end{aligned}$$

□

Proof of Theorem: Note that $\pi(k_x^n) = 1$ and so $||k_x^n - 1||_2^2 = \mathrm{Var}_\pi(k_x^n)$. Also, it is a fact that $K^*k_x^n = k_x^{n+1}$. And so when we apply Mihail's identity to k_x^n we obtain

$$||k_x^n - 1||_2^2 = ||k_x^{n+1} - 1||_2^2 + \varepsilon_M(k_x^n, k_x^n).$$

Then, since $\varepsilon_M(k_x^n, k_x^n) = \varepsilon_M(k_x^n - 1, k_x^n - 1)$, we may use the characterisation of the second largest eigenvalue of a π-reversible matrix given at (3) to give the inequality

$$||k_x^n - 1||_2^2 \geq ||k_x^{n+1} - 1||_2^2 + (1 - \beta_1(M))||k_x^n - 1||_2^2.$$

Induction then gives

$$||k_x^n - 1||_2^2 \leq \beta_1(M)^n ||k_x^0 - 1||_2^2$$

and the result follows from (7).

□

Remarks:

1. If K is reversible, then the eigenvalues of M are equal to $(\beta_i^2)_{i=0}^{N-1}$ where $(\beta_i)_{i=0}^{N-1}$ are the eigenvalues of K. Hence, $\beta_\star^2 = \beta_1(M)$ and so this result recovers the rate of convergence established before.

2. A corollary of this result is that if K is *strongly aperiodic* ($K(x,x) \geq \frac{1}{2}, \forall x$), then the same conclusion holds when $\beta_1(M)$ is replaced by $\beta_1(A)$. We show this by noting that the identity

$$M(K) \equiv A(K) + \frac{1}{4}M(2K - I) - \frac{1}{4}I$$

implies that, because $(2K - I)$ is a stochastic matrix,

$$\beta_1(M) \leq \beta_1(A) - \frac{1}{4}\{1 - \beta_1(M(2K - I))\}.$$

To see this, use the characterisation of eigenvalues given at (3). It is a fact that the eigenvalues of M are in $[0,1]$ and so

$$\beta_1(M) \leq \beta_1(A).$$

2.4 Eigenvalue bounds as $N \to \infty$

The results established so far hold for any Markov chain with a finite state space. In applications it is often the case, when we have family of related chains with increasing state space size, that we find

$$\beta_1(M) = e^{-1/f_1(N)}, \ \max_x\{1/\pi(x)\} = e^{f_2(N)}$$

with $f_i(N) \to \infty$ for $i = 1, 2$.

The bounds proved give us that $\max_x \|k_x^n - 1\|_2 \leq e^{-c}$ for $n \geq f_1(N)f_2(N) + 2cf_1(N)$. This means that the *time to stationarity*, T_2, of the chain is less than $f_1(N)(f_2(N)+2)$, where

$$T_2 := \min\{n > 0 : \ \max_x \|k_x^n - 1\|_2 \leq e^{-1}\}$$

and the subscript refers to the fact that we are measuring the distance with the l^2 norm. Let T_{TV} be the corresponding quantity when the distance used in the definition is total variation.

To illustrate this and the efficiency of eigenvalue bounds we discuss two examples:

Random walk on $\{0, 1, ..., N\}$

Define $S = \{0, 1, ..., N\}$ and let

$$K(x,y) = \begin{cases} \frac{1}{2} & \text{if } (x,y) = (0,0), (N,N) \text{ or } |x-y| = 1; \\ 0 & \text{otherwise.} \end{cases}$$

This defines an aperiodic, reversible Markov chain on S with invariant distribution $\pi(x) = \frac{1}{N+1}$. The eigenvalues of K are

$$\beta_j = \cos\left(\frac{\pi j}{N+1}\right), \ j = 0, 1, ..., N.$$

Thus, in the notation used above, we have that

$$f_1(N) = -\frac{1}{\ln\left(\cos\left(\frac{\pi}{N+1}\right)\right)}, \ f_2(N) = \ln N.$$

Hence, we have that $f_1(N)f_2(N) = O(N^2 \ln N)$ steps are sufficient to bring the chain close to stationarity. It turns out that only $O(N^2)$ steps are needed and so this bound is actually out by a factor $\ln N$.

Random walk on a group (continued)

Define our Markov chain as in section 2.2 with $P = (1-\theta)\delta_{g_0} + \theta u$. The chain is reversible and so we have that

$$\beta_1(M) = (1-\theta)^2$$

Thus, in the notation used above

$$f_1(|G|) = -\frac{1}{2\ln(1-\theta)},\ f_2(|G|) = \ln|G|.$$

Note that in this case $f_1(N) \not\to \infty$. However, the same argument applies and we have that

$$T_2 \leq -\frac{2+\ln|G|}{2\ln(1-\theta)}.$$

Furthermore, it is not difficult to show that

$$||k_{g_0}^n - 1||_2 = (|G|-1)^{1/2}(1-\theta)^n$$

which means that we may calculate explicitly that

$$T_2 = -\frac{2+\ln(|G|-1)}{2\ln(1-\theta)}$$

and so the upper bound for T_2 is particularly good. The same eigenvalue argument will not give a good bound on T_{TV} since the upper bound on $||K^n(g_0,\cdot) - \pi||_{TV}$ is poor.

3 Continuous time chains

In this section we look at continuous time chains and see how the convergence to equilibrium compares with the corresponding discrete time chains. We shall denote the continuous time semigroup associated with a Markov operator K by $(H_t)_{t\geq 0}$ and define it by $H_t = e^{-t(I-K)}$, i.e.

$$H_t^x(y) = H_t(x,y) = e^{-t}\sum_{n=0}^{\infty}\frac{t^n K^n(x,y)}{n}.$$

We note that when dealing with continuous chains we may drop the assumption that K is aperiodic.

The standard convergence result for continuous time Markov chains is

$$H_t(x,y) \to \pi(y) \text{ as } t \to \infty,\ \forall x,y \in S$$

where π is the invariant distribution defined by $\pi K = \pi$. We also have that π is invariant for H_t. Our main interest is in $||H_t^x - \pi||_{TV}$, but for technical reasons we introduce the density h_t^x which plays the role of k_x^n in continuous time. We shall define h_t^x by

$$h_t^x(y) = \frac{H_t^x(y)}{\pi(y)}$$

and note the inequalities

$$2||H_t^x - \pi||_{TV} \leq ||h_t^x - 1||_1 \leq ||h_t^x - 1||_2.$$

These allow the results we obtain for the l^2 distance to be used in bounding the variation distance. Finally, we introduce the adjoint operators $(H_t^*)_{t\geq 0}$. These form a Markov semigroup which satisfies $H_t^* = e^{-t(I-K^*)}$ and

$$\pi(x)H_t(x,y) = \pi(y)H_t^*(y,x). \tag{10}$$

3.1 Reversible continuous chains

In discrete time we saw how, when the chain is reversible, there is a particularly useful way in which to represent k_x^n in terms of the eigenvalues and eigenfunctions of K. Analogously, we may deduce a similar result for h_t^x, though now it turns out that it is more concise notationally to work with the matrix $I - K$.

Let K be reversible. Let $(\beta_i)_{i=0}^{N-1}$ and $(\psi_i)_{i=0}^{N-1}$ be the eigenvalues and eigenfunctions of K as introduced in section 2.2. If we set $\lambda_i = 1 - \beta_i$ then we have that $(\lambda_i)_{i=0}^{N-1}$ and $(\psi_i)_{i=0}^{N-1}$ are the eigenvalues and eigenfunctions of $I - K$ with $0 = \lambda_0 < \lambda_1 \leq ... \leq \lambda_{N-1}$. We may then deduce that

$$h_t^x(y) = \sum_{i=0}^{N-1} e^{-\lambda_i t}\psi_i(x)\psi_i(y)$$

by a similar argument to the one employed at (5). From this result, it is straightforward to obtain

$$||h_t^x - 1||_2^2 = \sum_{i=1}^{N-1} e^{-2t\lambda_i}\psi_i^2(x)$$

and consequently deduce bounds in terms of the λ_is. In particular, the bound corresponding to (8) is

$$||h_t^x - 1||_2^2 \leq \frac{1}{\pi(x)}e^{-2\lambda_1 t}. \tag{11}$$

Note how in the continuous case, the only eigenvalue of K that enters this bound is $\beta_1 = 1 - \lambda_1$. In the analogous bound for discrete time we needed to know the eigenvalue of largest modulus, $\beta_\star = \max\{\beta_1, |\beta_{N-1}|\}$.

Another observation that can be made is that if K is reversible then the convergence of the discrete and continuous chains are necessarily not too dissimilar. To qualify this we state the following bounds, from [12], which may be proved using fairly elementary methods.

1.
$$\|h_t^x - 1\|_2^2 \leq \frac{1}{\pi(x)} e^{-t} + \|k_x^{\lfloor t/2 \rfloor} - 1\|_2^2$$

2.
$$\|k_x^n - 1\|_2^2 \leq \beta_-^{2m}(1 + \|h_l^x - 1\|_2^2) + \|h_n^x - 1\|_2^2, \text{ for } n = m + l + 1$$

where $\beta_- = \max\{0, -\beta_{N-1}\}$. Taking $m = n - 1$ in this gives
$$\|k_x^n - 1\|_2^2 \leq \frac{1}{\pi(x)} \beta_-^{2(n-1)} + \|h_n^x - 1\|_2^2.$$

However, this is a result that cannot be extended to the irreversible case. A counterexample, as discussed by Saloff-Coste in [12], is a Markov chain on the integers, mod N, with $N = m^2$ an odd integer and
$$K(x, y) = \begin{cases} 1/2 & \text{if } y = x + 1, \\ 1/2 & \text{if } y = x + m \end{cases}$$

The discrete time chain can be shown to take $O(N^2)$ steps to be close to stationarity whereas the continuous time chain version takes only a time of $O(N)$.

3.2 Spectral gap

In discrete time, for irreversible K, we have seen how the distance to stationarity may be bounded by eigenvalues of a suitable reversibilization of K. There was a particularly straightforward way of doing this using the multiplicative reversibilization, $M(K)$. It turns out that in continuous time it is the additive reversibilization, $A(K)$, that is easier to work with.

The *spectral gap*, $\lambda = \lambda(K)$, is defined by
$$\lambda = \min\left\{\frac{\varepsilon_{A(K)}(f, f)}{\mathrm{Var}_\pi(f)} : \mathrm{Var}_\pi(f) \neq 0\right\}$$

and we also have

$$\lambda = \min\left\{\frac{\varepsilon_{A(K)}(f,f)}{\|f\|_2^2} \colon \pi(f) = 0\right\}.$$

We note from (3) that λ is the smallest non-zero eigenvalue of $I - A(K)$.

Theorem

$$\|h_t^x - 1\|_2^2 \leq \frac{1}{\pi(x)} e^{-2\lambda t}$$

To prove this we use the following lemma:

Lemma

$$\|H_t f - \pi(f)\|_2^2 \leq e^{-2\lambda t}\mathrm{Var}_\pi(f), \ \forall f \in l^2(\pi). \tag{12}$$

Proof of Lemma: Let $u(t) = \mathrm{Var}_\pi(H_t f)$. Then

$$u(t) = \|H_t f - \pi(H_t f)\|_2^2 = \|H_t f - \pi(f)\|_2^2$$

and also

$$u(t) = \langle H_t f, H_t f\rangle - \pi(H_t f)^2 = \langle H_t f, H_t f\rangle - \pi(f)^2.$$

Hence,

$$\begin{aligned} u'(t) &= \langle -(I-K)H_t f, H_t f\rangle + \langle H_t f, -(I-K)H_t f\rangle \\ &= -2\varepsilon_{A(K)}(H_t f, H_t f) \\ &\leq -2\lambda \mathrm{Var}_\pi(H_t f) = -2\lambda u(t). \end{aligned}$$

Thus,

$$\|H_t f - \pi(f)\|_2^2 = u(t) \leq e^{-2\lambda t}u(0) = e^{-2\lambda t}\mathrm{Var}_\pi(H_0 f) = e^{-2\lambda t}\mathrm{Var}_\pi(f).$$

□

Proof of Theorem: Let $f(y) = \mathbf{1}_x(y)/\pi(x)$. Using (10) we obtain

$$h_t^x(y) = \frac{\pi(x)H_t(x,y)}{\pi(y)}\frac{1}{\pi(x)} = H_t^*(y,x)\frac{1}{\pi(x)} = H_t^* f(y).$$

We have noted that $H_t^* = e^{-t(I-K^*)}$ defines a Markov semigroup. It follows from the definition of $A(K)$ that it has the same spectral gap as H_t and so the lemma implies that

$$\|h_t^x - \pi(f)\|_2^2 = \|H_t^* f - \pi(f)\|_2^2 \leq e^{-2\lambda t}\mathrm{Var}_\pi(f). \tag{13}$$

We also have that

$$\pi(f) = 1, \; \mathrm{Var}_\pi(f) = \frac{1-\pi(x)}{\pi(x)} \leq \frac{1}{\pi(x)}$$

and the result follows on substituting these into (13).

□

Remark: We note that when K is reversible then $\lambda(K) = \lambda_1$ and so (11) is a special case of this result.

4 Geometric bounds on eigenvalues

We have now seen that the second largest eigenvalue of a reversible Markov operator is of particular interest when analysing the rate of convergence to equilibrium. The relevant eigenvalues being β_1 in the reversible discrete case, $\beta_1(M)$ in the general discrete case and $\beta_1(A) \equiv 1 - \lambda$ in the continuous case.

In this section we assume that K is reversible and discuss geometrical methods that can be used to bound β_1. This is equivalent to finding bounds on the eigenvalue $\beta_1(M)$ and λ in the general case because $M(K)$ and $A(K)$ are both reversible Markov operators. We note that, for reversible K, $A(K) = K$ and so $\lambda = 1 - \beta_1$.

We also present a lower bound on the minimum eigenvalue for a reversible chain, which allows us to construct a bound on $\beta_\star = \min\{\beta_1, |\beta_{N-1}|\}$, which is also relevant to the topic of convergence in the discrete case.

The bounds on the eigenvalues are given in terms of quantities that depend on path counting, allowing us to approach what is essentially a linear mathematics problem in a combinatorial manner.

4.1 Conductance

We first define the probability measure, Q, on $S \times S$ by

$$Q(x,y) = \pi(x)K(x,y)$$

and then define the *conductance* to be

$$\Phi = \min_{A \subseteq S, 0 < \pi(A) \leq \frac{1}{2}} \frac{Q(\partial A)}{\pi(A)}.$$

where $Q(\partial A) := \sum_{x \in A, y \in A^c} Q(x,y)$. The following result shows how this parameter can be used to bound the spectral gap $\lambda = 1 - \beta_1$, the proof presented here is due to Diaconis/Stroock [7].

Theorem (Cheeger's inequality)

$$\frac{1}{2}\Phi^2 \leq \lambda \leq 2\Phi.$$

In fact, the proof of the theorem gives a slightly better upper bound using

$$\Phi' = \min_{A \subseteq S} \frac{Q(\partial A)}{2\pi(A)(1-\pi(A))}$$

which clearly satisfies $\Phi' \leq \Phi$.

Outline of proof: The upper bound is relatively straightforward to prove and relies on setting $f = 1_B$, for $B \subseteq S$ in the definition of the spectral gap. It is not difficult to calculate that

$$\varepsilon_K(1_B, 1_B) = \frac{1}{2}\sum_{x,y}(1_B(x) - 1_B(y))^2 K(x,y)\pi(x) = Q(\partial B).$$

Which gives $\forall B \subseteq S$,

$$\lambda = \min\left\{\frac{\varepsilon_K(f,f)}{\mathrm{Var}_\pi(f)} : \mathrm{Var}_\pi(f) \neq 0\right\} \leq \frac{\varepsilon_K(1_B, 1_B)}{\mathrm{Var}_\pi(1_B)} = \frac{Q(\partial B)}{\pi(B)(1-\pi(B))}.$$

Comparing this to the the definition of Φ' yields the bound $\lambda \leq 2\Phi'$.

The lower bound is not a triviality and does require some thought. Diaconis and Stroock prove the bound with two main steps of which the main ideas are as follows.

Step 1: Proof that if $f_+ = f \vee 0$ and $S(f) = \{x\colon f(x) > 0\}$ then $\forall f$ with $S(f) \neq \emptyset$ and $k \in [0, \infty)$ we have

$$k\pi(f_+^2) \geq \varepsilon_K(f_+, f_+) \tag{14}$$

if $(I - K)f \leq kf$ on $S(f)$.

This can proved by merely expanding out the inner product and noticing that $(f_+(x) - f_+(y))^2 \leq (f_+(x) - f_+(y))(f(x) - f(y))$.

Step 2: Proof that for a positive function, f, we have

$$2\varepsilon(f,f)_K \geq \Phi(f)^2\pi(f^2) \tag{15}$$

where $\Phi(f) := \min\{Q(\partial A)/\pi(A) :\ \emptyset \neq A \subseteq S(f)\}$.

Firstly, by an application of Cauchy-Schwarz and substituting for Q we obtain the inequality

$$\sum_{x,y} |f(x)^2 - f(y)^2| Q(x,y) \leq \sqrt{8\pi(f^2)\varepsilon_K(f,f)}. \tag{16}$$

Perhaps the most ingenious part of the proof is in the following calculation where $Q(\partial A_t)$ is introduced and allows a comparison with $\Phi(f)$.

$$\begin{aligned}
\sum_{x,y} |f(x)^2 - f(y)^2| Q(x,y) &= 4 \sum_{f(x)<f(y)} \int_{f(x)}^{f(y)} t dt Q(x,y) \\
&= 4 \int_0^\infty t \sum_{f(x)\leq t<f(y)} Q(x,y) dt \\
&= 4 \int_0^\infty t Q(\partial A_t) dt, \text{ where } A_t = \{x\colon\ f(x) > t\} \\
&\geq 4\Phi(f) \int_0^\infty t\pi\{x\colon\ f(x) > t\} dt \\
&= 2\Phi(f)\pi(f^2)
\end{aligned}$$

Combining this with (16) completes step 2.

Completing the proof is now a matter of putting (14) and (15) together to obtain

$$k \geq \frac{\Phi(f_+)^2}{2}, \text{ if } (I - K)f \leq kf \text{ on } S(f).$$

Now choose f to be the eigenfunction of $(I - K)$ with eigenvalue λ so that the condition is satisfied everywhere with $k = \lambda$. Since $\pi(f) = 0$, f can be chosen so that $\pi(S(f_+)) = \pi(S(f)) \leq \frac{1}{2}$. Hence, we have that $\Phi(f_+) \geq \Phi$ which completes the proof.

□

4.2 Poincaré inequality

A *Poincaré inequality* is of the form

$$\mathrm{Var}_\pi(f) \leq C\varepsilon_K(f, f), \forall f.$$

As noted in [12], the definition of the spectral gap implies that if we have a C that satisfies a Poincaré inequality then necessarily $1/C \leq \lambda$. We will see various combinatorial quantities that satisfy such an inequality and which can consequently be used to bound λ.

First, define Σ to be a collection of adapted paths, γ_{xy}, from x to y (one for each pair (x, y)). Here, we say a path is *adapted* if for each edge, $e = (u, v)$, on the path we have $K(u, v) > 0$, or equivalently if $Q(e) > 0$. Sinclair, [13], then defines $\overline{\rho}$ by

$$\overline{\rho} = \overline{\rho}(\Sigma) = \max_{e:Q(e)\neq 0} \left\{ \frac{1}{Q(e)} \sum_{x,y\in S: e\in\gamma_{xy}} |\gamma_{xy}|\pi(x)\pi(y) \right\},$$

where $|\gamma_{xy}|$ is the length of the path γ_{xy}.

Theorem

$$\lambda \geq \frac{1}{\overline{\rho}}$$

Proof: As indicated above, the result is proved on showing that $\overline{\rho}$ satisfies a Poincaré inequality. Note that, by Cauchy-Schwarz

$$|f(y) - f(x)|^2 \leq \left(\sum_{(u,v)\in\gamma_{xy}} |f(v) - f(u)| \right)^2 \leq |\gamma_{xy}| \sum_{(u,v)\in\gamma_{xy}} |f(v) - f(u)|^2$$

Hence,

$$\begin{aligned}
\mathrm{Var}_\pi(f) &= \frac{1}{2}\sum_{x,y} |f(y) - f(x)|^2 \pi(x)\pi(y) \\
&\leq \frac{1}{2}\sum_{x,y}\sum_{(u,v)\in\gamma_{xy}} |f(v) - f(u)|^2 |\gamma_{xy}|\pi(x)\pi(y) \\
&= \frac{1}{2}\sum_{u,v} |f(v) - f(u)|^2 \left\{ \frac{1}{Q(u,v)} \sum_{(x,y):(u,v)\in\gamma_{xy}} |\gamma_{xy}|\pi(x)\pi(y) \right\} Q(u,v) \\
&\leq \frac{\overline{\rho}}{2}\sum_{u,v} |f(v) - f(u)|^2 Q(u,v) \\
&= \varepsilon_K(f,f)\overline{\rho}
\end{aligned}$$

□

Sinclair also defines

$$\rho = \max_{e:Q(e)\neq 0} \left\{ \frac{1}{Q(e)} \sum_{\gamma_{xy}:e\in\gamma_{xy}} \pi(x)\pi(y) \right\}$$

which is clearly easier to calculate than $\overline{\rho}$ because it no longer depends on the lengths of the paths. We note that if $l := \max_{\gamma_{xy}\in\Sigma} |\gamma_{xy}|$, the length of the longest path, then $\overline{\rho} \leq \rho l$. Which gives the lower bound for λ of $\frac{1}{\rho l}$.

We also note without proof the following proposition (simple proofs of the result appear in [7], [12] and [13]).

Proposition

$$\Phi \geq \frac{1}{2\rho}$$

This gives, in conjunction with Cheeger's inequality, that $\frac{1}{8\rho^2} \leq \lambda$. Sinclair argues that frequently we will have $l \ll \rho$ and so this bound will not be as sharp as $\frac{1}{\rho l}$. Further discussion of comparisons between Cheeger and Poincaré type inequalities appears in section 4.4.

A similar quantity satisfying a Poincaré inequality is defined in [7] to be

$$\kappa = \max_e \left\{ \sum_{\gamma_{xy}:e\in\gamma_{xy}} |\gamma_{xy}|_Q \pi(x)\pi(y) \right\}, \text{ where } |\gamma_{xy}|_Q := \sum_{e\in\gamma_{xy}} Q(e)^{-1}.$$

In the case of random walks on graphs, it is remarked by Sinclair, that $\overline{\rho}$ and κ coincide and so the bounds provided by them are both identical. Conversely, in some instances, such as the Ehrenfest Urn model, $\overline{\rho}$ vastly outperforms κ. In general, as remarked in both [7] and [13], the bounds are incomparable.

It should be noted that these bounds are by no means guaranteed to be optimal. Sinclair generalises both ρ and $\overline{\rho}$ to depend on a "flow function" defined on the edges of a weighted graph. After doing this, he derives similar looking bounds on λ to those already described and presents examples in which these flow dependent parameters outperform the ones that have been defined here.

4.3 A bound on the minimum eigenvalue

Another useful bound obtained in [7] is for the minimum eigenvalue, β_{N-1}, of a reversible Markov chain. We first define Σ' to be a collection of odd length adapted paths, σ_x, from x to x (one for each x) and define $|\sigma_x|_Q$ analogously to $|\gamma_{xy}|_Q$. Now, define the geometric quantity

$$\iota = \iota(\Sigma') = \max_e \left\{ \sum_{\sigma_x:e\in\sigma_x} |\sigma_x|_Q \pi(x) \right\}.$$

This can be used to bound β_{N-1} as follows.

Proposition

For reversible K,

$$\beta_{N-1} \geq -1 + \frac{2}{\iota}.$$

Proof: The reason for only considering odd paths becomes apparent in the proof for it allows us to represent $f(x)$ as $\frac{1}{2}((f(x)+f(y))-(f(y)+f(w))+...+$

$(f(z) + f(x))$. For the purposes of the following calculation only, introduce the notation $e = (e_-, e_+)$ and let $l(e)$ be the distance of e_- from x in σ_x.

$$\begin{aligned}
\pi(f^2) &= \frac{1}{4}\sum_x \pi(x)\left\{\sum_{e\in\sigma_x}(-1)^{l(e)}(f(e_-)+f(e_+))\right\}^2 \\
&\leq \frac{1}{4}\sum_x \pi(x)\left(\sum_{e\in\sigma_x}(f(e_-)+f(e_+))^2Q(e)\right)\left(\sum_{e\in\sigma_x}Q(e)^{-1}\right) \\
&= \frac{1}{4}\sum_e (f(e_-)+f(e_+))^2Q(e)\sum_{\sigma_x:e\in\sigma_x}|\sigma_x|_Q\pi(x) \\
&\leq \frac{\iota}{4}\sum_e (f(e_-)+f(e_+))^2Q(e) \\
&= \frac{\iota}{2}\left(\pi(f^2)+\langle f, Kf\rangle\right)
\end{aligned}$$

Here, the first inequality is Cauchy-Schwarz. The final equality is actually an identity. The result follows on applying the characterisation of the eigenvalues of a reversible Markov operator given at (4).

□

Diaconis/Stroock remark that this result can be adapted to allow the paths considered to be σ_{xy}, odd length paths from x to y for each pair (x, y). Correspondingly, let

$$\iota' = \max_e\left\{\sum_{\sigma_{xy}:e\in\sigma_{xy}}|\sigma_{xy}|_Q\pi(x)\pi(y)\right\}.$$

We then obtain $\beta_{N-1} \geq -1 + 1/\iota'$ and since the paths σ_{xy} can also be used in the definition of κ we obtain that

$$\beta_\star \leq 1 - \frac{1}{\iota'}.$$

4.4 Comparison of Cheeger and Poincaré bounds

A question that remains unanswered in general is, for which situations do Cheeger type inequalities provide better bounds than Poincaré inequalities? If we define on a finite graph E to be the edge set, $d_\star$ to be the maximum degree, $\gamma_\star$ to be the maximum path length (where there is one path chosen for each ordered pair of vertices) and $b = \max_e |\{\gamma_{xy} : e \in \gamma_{xy}\}$ to be the "maximum bottlenecking" then we obtain easily that

$$\rho \leq \frac{bd_\star^2}{2|E|} \text{ and } \overline{\rho} \leq \frac{bd_\star^2\gamma_\star}{2|E|}$$

for a simple random walk on the graph. And hence obtain a Cheeger bound of

$$\frac{1}{2}\left(\frac{|E|}{d_\star^2 b}\right)^2 \leq \lambda \tag{17}$$

and a Poincaré bound of

$$\frac{2|E|}{d_\star^2 \gamma_\star b} \leq \lambda. \tag{18}$$

In a paper by Fulman/Wilmer, [10], it is demonstrated that (18) is superior to (17) for simple random walks on trees and vertex transitive graphs. The paper also shows how for weighted random walks on a vertex transitive graph with a uniform stationary distribution, the bound $(8\rho^2)^{-1}$ is worse than $\overline{\rho}^{-1}$.

Conversely, in [13], Sinclair demonstrates that for an asymmetric random walk on $\{0, ..., N-1\}$ that $(8\rho^2)^{-1}$ differs from the true value asymptotically only by a constant factor and that this turns out to be asymptotically much better than the estimate provided by $\overline{\rho}^{-1}$. This removes any chance of extending Fulman/Wilmer's results to weighted trees.

A final point of note is that geometric bounds are not always the best way of proceeding. An example illustrating this, noted in [10], is of a random walk on a Ramanujan graph. This is a p regular graph on $\frac{1}{2}q(q^2-1)$ vertices where $p, q \equiv 1 \pmod 4$. The first eigenvalue of the chain is known and $\lambda = 1 - \frac{2\sqrt{p}}{p+1} > 0$. It is explained that the best path-based bounds will not be better than $8/(\log_p \frac{1}{2}q(q^2-1))^2$ which tends to 0 as $q \to \infty$. However, using number theoretic methods to bound Φ directly, a lower bound for λ that is greater than 0, independent of q, may be obtained.

5 The cutoff phenomenon

So far, we have concentrated on upper bounds on the distance from stationarity. This allows us to make comments on how long is sufficient to run a chain so that it is close to equilibrium. The upper bounds we have seen are of an exponential type and so, at worst, the convergence will happen at this rate. As it turns out, when we do analyse the convergence more carefully, we see in some families of chains that, asymptotically, the convergence is a lot more sudden.

There are many ways to define what is known as the *cutoff phenomenon* precisely, but the basic idea is that the chain will stay a long way from stationarity up to some time before rapidly approaching equilibrium. We present here discrete time versions of definitions set out by Saloff-Coste in [12] which formalise the cutoff point for a family of chains.

First, let $\mathcal{F} = \{(S_N, K_N, \pi_N) : N = 1, 2, ...\}$ be an infinite family of finite chains. Then

1. one says that $\mathcal{F}$ presents a *cutoff in total variation* or *variation cutoff* with critical time $(t_N)_1^\infty$ if $t_N \to \infty$ and $\forall \varepsilon > 0$ we have

$$\lim_{N\to\infty} \max_{x\in S_N} \|K_N^{\lfloor t_N(1-\varepsilon)\rfloor}(x, \cdot) - \pi_N\|_{TV} = 1$$

   and

$$\lim_{N\to\infty} \max_{x\in S_N} \|K_N^{\lfloor t_N(1+\varepsilon)\rfloor}(x, \cdot) - \pi_N\|_{TV} = 0.$$

2. if $(t_N, b_N)_1^\infty$ are such that $t_N, b_N \geq 0$, $t_N \to \infty$ and $b_N/t_N \to 0$, one says that $\mathcal{F}$ presents a *cutoff of type* $(t_N, b_N)_1^\infty$ *in total variation* if, for all real c, we have

$$\lim_{N\to\infty} \max_{x\in S_N} \|K_N^{\lfloor t_N+b_N c\rfloor}(x, \cdot) - \pi_N\|_{TV} = f(c),$$

   with $f(c) \to 1$ as $c \to -\infty$ and $f(c) \to 0$ as $c \to \infty$.

The second definition here clearly implies the first and in fact gives a slightly more precise description of what is happening near to the cutoff point. As is noted in [12], these definitions mean that in order to approximate π_N one should not stop the chain until t_N but it is essentially useless to run the chain for longer. There are many variations on this definition. For example, we can also define a cutoff relative to other distances or in terms of continuous time.

A similar property to the cutoff phenomenon is introduced by Diaconis/Aldous, [2]. They consider each Markov chain started from a fixed initial position, $x_0^N \in S_N$ say, and then consider the distance from stationarity of the distribution of the chain after n steps. The corresponding effect they call the *threshold phenomenon*. Indeed, they define a *variation threshold* in the same way as we have defined a variation cutoff but with the $\max_{x\in S}$ removed. During the course of this essay we shall see both an example in which the threshold phenomenon occurs (5.1) and one in which there is no cutoff (6.2).

It is evident that a threshold is a weaker property than a cutoff since a cutoff requires some kind of uniformity of convergence over the whole of S_N. One example of where the start points $(x_0^N)_{N=1}^\infty$ make a difference is noted by Diaconis, [4]. The example in question is the Ehrenfest urn model with N balls. Let X_n^N be the number of balls in a particular urn after n picks. To remove the periodicity problem we give some probability to no ball moving and define K_N by

$$K_N(x, x-1) = \frac{x}{N+1}, \; K_N(x,x) = \frac{1}{N+1}, \; K_N(x, x+1) = \frac{N-x}{N+1}$$

for appropriate x. Then if $X_0^N \equiv 0$, the chain has a variation threshold at $\frac{1}{4}N \ln N$, whereas for $X_0^N \equiv N/2$ there is no such threshold.

It is also perhaps interesting to note that in the case $X_0^N \equiv N/2$, only $O(N)$ steps are necessary suffice to achieve stationarity. This observation could be relevant in practical situations as it may be advantageous to choose the start point of the chain so as to reduce the convergence time.

5.1 Product random walk threshold

In this section we present an example from [2] showing how we can infer a separation threshold for a product random walk given that we know something about the convergence of a single random walk on a group. We also quote the result for the corresponding variation threshold.

For each N, define $(X_n^N)_{n=0}^\infty$ to be the random walk on G^N corresponding to the distribution P^N, i.e. $(X_n^N)_{n=0}^\infty$ is a Markov chain on G^N, starting from the identity, g_0^N, with transition matrix K_N defined by

$$K_N(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{N} P(y_i x_i^{-1}).$$

Let $K = K_1$. We note that if K is not irreducible or aperiodic (equivalently P is not supported on any coset of a proper subgroup of G), the equilibrium distribution π_N is uniform on G^N.

We assume further that for each $x \in G$

$$K^n(g_0, x) - |G|^{-1} \sim a_x e^{-n/\tau}. \qquad (19)$$

This is not a ridiculous assumption since, as suggested by the form of K^n given at (5), we typically find $K^n(x,y) - \pi(y) \sim c_{x,y}|\beta_1|^n$ for some suitable constants $c_{x,y}$.

Theorem

Under assumption (19), $\mathcal{F} = \{(G^N, K_N(g_0^N, \cdot), \pi_N)\colon N = 1, 2, ...\}$ has a separation threshold with critical time $(s_N)_1^\infty$ where

$$s_N := \tau \ln N.$$

Also, $\mathcal{F}$ has a variation threshold with critical time $(t_N)_1^\infty$ where

$$t_N := \frac{\tau}{2} \ln N.$$

Proof of separation threshold: It is clear from the definition of K_N that

$$K_N^n(\mathbf{x}, \mathbf{y}) = \prod_{i=0}^{N} K^n(x_i, y_i).$$

Hence,

$$\begin{aligned} d_{sep}(K_N^n(g_0^N, \cdot), \pi_N) &= \max_{\mathbf{x} \in G^N} \left\{1 - |G|^N K_N^n(g_0^N, \mathbf{x})\right\} \\ &= 1 - \left\{|G \min_{x \in G} K^n(g_0, x)\right\}^N \\ &= 1 - \{1 - d_{sep}(K^n(g_0, \cdot), \pi_1)\}^N \end{aligned}$$

Assumption (19) gives us that

$$d_{sep}(K^n(g_0, \cdot), \pi_1) \sim Ae^{-n/\tau}$$

where $A := \max(-a_x)$ and so

$$d_{sep}(K_N^n(g_0^N, \cdot), \pi_N) \sim 1 - (1 - Ae^{-n/\tau})^N,$$

from which it follows that for $\varepsilon > 0$,

$$\begin{aligned} d_{sep}(K_N^{\lfloor \tau \ln N(1-\varepsilon) \rfloor}(g_0^N, \cdot), \pi_N) &\to 1, \text{ as } N \to \infty \\ d_{sep}(K_N^{\lfloor \tau \ln N(1+\varepsilon) \rfloor}(g_0^N, \cdot), \pi_N) &\to 0, \text{ as } N \to \infty \end{aligned}$$

and so $\mathcal{F}$ has the desired threshold.

□

5.2 Weak cutoff

Although the proof of the separation threshold on the product group is quite straightforward under the assumptions that are used, in many situations verifying that the conditions of the definition of a threshold/cutoff are satisfied is not a trivial matter. This leads to the formulation of a weaker definition that is easier to work with. The definition quoted here is the l^2 version of the definition given in [12].

Let $\mathcal{F} = \{(S_N, K_N, \pi_N) : N = 1, 2, ...\}$ be as above. Set $H_{N,t} = e^{-t(I-K_N)}$ and then

1. one says that $\mathcal{F}$ presents a *weak l^2-cutoff* with critical time $(t_N)_1^\infty$ if $t_N \to \infty$ and $\forall \varepsilon > 0$ we have
$$\lim_{N\to\infty} \max_{x\in S_N} \|h^x_{N,t_N(1-\varepsilon)} - 1\|_2 > 0$$
and
$$\lim_{N\to\infty} \max_{x\in S_N} \|h^x_{N,t_N(1+\varepsilon)} - 1\|_2 = 0.$$

2. if $(t_N, b_N)_1^\infty$ are such that $t_N, b_N \geq 0$, $t_N \to \infty$ and $b_N/t_N \to 0$. One says that $\mathcal{F}$ presents a *weak l^2-cutoff of type* $(t_N, b_N)_1^\infty$ if, for all real c, we have
$$\lim_{N\to\infty} \max_{x\in S_N} \|h^x_{N,t_N+b_N c} - 1\|_2 = f(c),$$
with $f(0) > 0$ and $f(c) \to 0$ as $c \to \infty$.

This weaker definition allows us to give a sufficient condition for a weak cutoff in terms of the spectral gaps of the chains.

Theorem

Fix $\varepsilon > 0$. Let $\mathcal{F}$ be as previously defined. Let λ_N be the spectral gap of K_N and let
$$t_N = \min\{t > 0 : \max_{x\in S_N} \|h^x_{N,t} - 1\|_2 \leq \varepsilon\}.$$
A sufficient condition for $\mathcal{F}$ to present a weak l^2-cutoff with critical time $(t_N)_1^\infty$ is
$$\lim_{N\to\infty} \lambda_N t_N = \infty. \tag{20}$$
In fact, when this condition occurs we can deduce that $\mathcal{F}$ has an l^2-cutoff of type $(t_N, 1/\lambda_N)_1^\infty$. Also, if the chains are reversible then the converse holds.

Proof: First, assume that (20) holds. By the definition of t_N we have that $\max_{x\in S_N} \|h^x_{N,t_N} - 1\|_2 = \varepsilon > 0$. We also have that, $\forall x \in S_N$,
$$\begin{aligned}\|h^x_{N,t_N+s} - 1\|_2 &= \|(H^*_{N,s} - \pi_N)(h^x_{N,t_N} - 1)\|_2 \\ &\leq \sup\left\{\frac{\|(H^*_{N,s} - \pi_N)f\|_2}{\|f\|_2} : \|f\|_2 \neq 0\right\} \|h^x_{N,t_N} - 1\|_2 \\ &\leq e^{-s\lambda_N}\varepsilon\end{aligned}$$

Where the last inequality is a direct consequence of equation (12). Hence

$$\lim_{N\to\infty}\max_{x\in S_N}\|h^x_{N,t_N+s/\lambda_N}-1\|_2=f(s)$$

satisfies $f(s)\leq\varepsilon e^{-s}$ and $f(0)>0$ so that $\mathcal{F}$ has the desired cutoff.

Conversely, suppose that (20) does not hold and that each chain is reversible. Since (20) does not hold, there exists a subsequence with $\lambda_{N_i}t_{N_i}\leq a$ for some finite $a>0$. For a reversible chain, H_t, it is a fact that

$$\max_{x\in S}\|H^x_t-\pi\|_1\geq e^{-t\lambda}.$$

We also have that $\|H^x_{N,t}-\pi_N\|_1\leq\|h^x_{N,t}-1\|_2$. Thus, $\forall\varepsilon\geq 0$,

$$\max_{x\in S_{N_i}}\|h^x_{N_i,t_{N_i}(1+\varepsilon)}-1\|_2\geq e^{-t_{N_i}\lambda_{N_i}(1+\varepsilon)}\geq e^{-a(1+\varepsilon)}>0$$

and consequently

$$\max_{x\in S_N}\|h^x_{N,t_N(1+\varepsilon)}-1\|_2\not\to 0$$

which means that there can be no weak l^2-cutoff.

□

In fact, the sufficiency of (20) for a weak cutoff does actually apply to other l^p distances, including l^∞, when the times t_N are suitably modified. The details of this are omitted here as they do not really throw any more light on the phenomenon and the proofs require only minor extensions of that which has been presented.

6 Applications

We have now discussed several areas related to determining the rate of convergence of Markov chains, as yet though we have not really considered why this may be of interest. It turns out that simulating a distribution, π, on a large state space by running a Markov chain with π as its invariant distribution is of great practical interest in areas such as statistical physics, statistics and computer science.

In these applications it is important to know when is a good time to stop the Markov chain. If we stop too early, the distribution we sample may not represent π well and conversely, due to the finite speed of the computers that will be running the simulations, it is also not possible to continue indefinitely. Hence, it is useful to be able to generate a sharp bound for the distance from stationarity of the simulated Markov chain to enable us to achieve this balance. Note also the particular advantage a cutoff gives. For, if a cutoff can be proved, this determines fairly precisely what sort of time will be both necessary and sufficient to allow a reasonable simulation.

The bounds given in earlier sections are not the only methods used to bound convergence. Often, when specific Markov chains are considered a better analysis may be carried out. For example, in the second application we present, we see how the form of S allows the use of Fourier analysis. There are also statistical methods to estimate parameters such as eigenvalues which can then be used to gauge the convergence time. For instance, in [9], Garren/Smith consider a method of estimating the second largest eigenvalue of the chain and use this to produce an estimate on how many steps will be necessary to bring the chain close to equilibrium. However, statistical methods such as these are not explored in this essay.

In this section we describe results from two main applications detailing different aspects of this area. The first involves material from [5] and we give examples in which the Metropolis algorithm's convergence rate may be sharply bounded by eigenvalue methods. The second application is the generation of random numbers on the integers mod N by Markov chain methods.

6.1 The Metropolis algorithm

Outline of the Metropolis algorithm

One widely used Markov chain Monte-Carlo (simulation) method is the Metropolis algorithm which allows us to generate a chain with invariant distribution π from the ratios $r_{yx} = \pi(y)/\pi(x)$ only. This is particularly relevant to large state space chains where calculating the normalising constant is infeasible.

Suppose we have a "base chain" on a finite state space, S. This is a symmetric, irreducible Markov chain that is easy to run. Let the transition matrix for this chain be K. We then define the stochastic matrix M by

$$M(x,y) = \begin{cases} K(x,y)r_{yx} & \text{if } r_{yx} < 1, \\ K(x,y) & \text{if } x \neq y \text{ and } r_{yx} \geq 1, \\ K(x,y) + \sum_{z \neq x, r_{zx} < 1} K(x,z)(1 - r_{zx}) & \text{if } x = y. \end{cases}$$

This defines an irreducible, aperiodic and reversible Markov transition matrix with invariant distribution π. Furthermore, the formulation also allows a simple implementation of the new chain using merely the base chain and coin tossing simulation (see [5] for a greater description). This setup means that it is the convergence of $M(x, \cdot)$ to π that is of interest.

Metropolis algorithm using a random walk on $\{0,1\}^N$

In this example, we see how a random walk on $\{0,1\}^N$ may be used to generate a Markov chain with invariant distribution, π_θ, defined by

$$\pi_\theta(x) = \frac{\theta^{H(0,x)}}{(1+\theta)^N}, \text{ for } x \in \{0,1\}^N$$

where $H(x,y)$ is the Hamming distance between x and y. We also show how eigenvalues of a Markov chain related to the Metropolis chain may be used to bound the convergence.

In this example we consider our base chain to be a random walk on $\{0,1\}^N$ with non-zero transition probabilities being $K(x,y) = \frac{1}{N}$ for $H(x,y) = 1$. Hence, we may use the prescription given above to write down M.

$$M(x,y) = \begin{cases} \frac{1}{N} & \text{if } H(x,y) = 1 \text{ and } H(0,y) < H(0,x), \\ \frac{\theta}{N} & \text{if } H(x,y) = 1 \text{ and } H(0,y) > H(0,x), \\ (1 - \frac{H(0,x)}{N})(1-\theta) & \text{if } H(x,y) = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Perhaps the most useful observation of the discussion is that for a permutation of "bits", $\sigma(x_1, ..., x_N) = (x_{\sigma(1)}, ..., x_{\sigma(N)})$, we have $M(\sigma x, \sigma y) = M(x,y)$. This means that the chain which records the weight of x, $H(0,x)$, is a Markov chain taking values in $\{0, ..., N\}$ and with transition probabilities

$$m(i,j) = \begin{cases} \frac{i}{N} & \text{if } j = i-1, \\ (1-\frac{i}{N})\theta & \text{if } j = i+1, \\ (1-\frac{i}{N})(1-\theta) & \text{if } j = i. \end{cases}$$

Let the invariant distribution for this chain be μ_θ. We have that, because of the permutation invariance,

$$||M^n(0,\cdot) - \pi_\theta||_{TV} = ||m^n(0,\cdot) - \mu_\theta||_{TV}.$$

Hence, an upper bound for $||M^n(0,\cdot) - \pi_\theta||_{TV}$ may be given in terms of the eigenvalues and eigenvectors for m which are known explicitly and are given in [5]. It turns out that

$$||M^n(0,\cdot) - \pi_\theta||_{TV} \leq f(\theta, c), \text{ for } n = \frac{N}{2(1+\theta)}(\ln N\theta + c) \tag{21}$$

where $f(\theta, c) \to 0$ for $c \to \infty$ independent of N. In fact, this example illustrates how useful eigenvalue estimates can be since, also noted in [5], is the result

$$\lim_{N \to \infty} \|M^n(0, \cdot) - \pi_\theta\|_{TV} > 0, \text{ for } n = \frac{N}{2(1+\theta)}(\ln N\theta - c)$$

and so we can not improve on (21) asymptotically. It perhaps ought to be noted that this result shows how the Metropolis algorithm is not optimal in simulating from π_θ since there do exist methods for doing so which are $O(N)$.

Simulation from a distribution on the symmetric group

Although the previous example is fairly simple, parts of the argument are well worth noting for use in more complicated chains. For example, consider the distribution on the symmetric group, S_N, given by

$$\pi_\theta(\sigma) = k(\theta)\theta^{c(\sigma, \sigma_0)},$$

where $k(\theta)$ is a normalising constant and $c(\sigma, \sigma_0)$ is the Cayley distance between σ and σ_0 (this is defined to be minimum number of transpositions needed to bring σ to σ_0). There is a practical reason for considering distributions such as this on S_N. For instance, statisticians may need to work with ranked data and this type of distribution may be helpful when analysing the data.

The base chain for the Metropolis algorithm is assumed to be a particular random walk on S_N with transition probabilities

$$K(\sigma, \tau) = \begin{cases} 1/\binom{N}{2} & \text{if } \tau = \sigma(i,j) \text{ for some } i < j, \\ 0 & \text{otherwise.} \end{cases}$$

where (i, j) is the transposition of i and j. When the transition matrix for the metropolis chain is constructed it transpires that, if σ and τ are conjugate in S_N, then $M(\text{id}, \sigma) = M(\text{id}, \tau)$. This means that, similarly to the previous example, we may map down to a chain on a smaller state space and consider the eigenanalysis of this. In this case, it is the Markov chain that records only the conjugacy classes of the Metropolis chain.

Full analysis of this example is provided in [5]. Again, the authors suggest that the eigenvalue bound does result in the optimal bound asymptotically.

6.2 Random number generation

Generating random numbers on a computer is a matter of great practical interest. One method for generating pseudo-random sequences from

$\{0, ..., N-1\}$ is to compute them recursively using formulae such as $X_{n+1} = aX_n + b \pmod{N}$. Although a and b will be chosen so that $(X_n)_0^\infty$ has some of the same properties as a random sequence, the method has the obvious problem of being deterministic.

In this section, we look at the related recursion

$$X_{n+1} = aX_n + b_n (\text{mod } N) \tag{22}$$

where the b_n are independent, uniform on $\{-1, 0, 1\}$. Recursions similar to this are sometimes used in computer graphics algorithms. The question that we shall be considering is how many steps will it take for the distribution of X_n to become close to uniform on $\{0, ..., N-1\}$. We look at the cases $a = 1$ and $a = 2$ and see how the deterministic doubling that occurs when $a = 2$ is, perhaps surprisingly, an effective way to reduce the time that the sequence takes to become close to random.

In the case $a = 1$ we have a simple random walk on the integers, mod N, with some probability given to not moving. The chain $(X_n)_0^\infty$, defined by the recursion and $X_0 = 0$, is clearly irreducible, aperiodic Markov with invariant distribution being uniform on $\{0, ..., N-1\}$. We shall show how the chain takes of the order of N^2 steps to reach stationarity and exhibits no cutoff. In doing this, we shall also demonstrate the usefulness of Fourier analysis techniques when dealing with problems on such a state space. The argument presented here follows that given in [3].

Throughout this section define $S_N := \{0, ..., N-1\}$ and define the *Fourier transform* of a function, f, on S_N to be

$$\hat{f}(x) = \sum_{y \in S_N} e_N^{xy} f(y)$$

where $e_N := e^{2\pi i/N}$. In the proof of the following theorem we shall use the identity

$$\sum_{x \in S_N} |\hat{f}(x)|^2 = N \sum_{x \in S_N} |f(x)|^2$$

and also the result that for two functions, f and g, we have:

$$(f * g)\hat{} = \hat{f}\hat{g}$$

where $f * g$ is the convolution of f and g defined by $f * g(x) = \sum_y f(y) g(y-x)$. Furthermore, let K_N, π_N be the transition matrix and invariant distribution, respectively, for the chain on S_N.

Theorem

There exist $\alpha, \beta > 0$ such that

$$e^{-\frac{\alpha n}{N^2}} \leq ||K_N^n(0,\cdot) - \pi_N||_{TV} \leq e^{-\frac{\beta n}{N^2}}. \tag{23}$$

Outline of proof: For clarity, in this proof set $K^n(j) = K_N^n(0,j)$. Using Cauchy-Schwarz and noting that $\hat{K}^n(0) = 1$ it may be verified that

$$||K^n - \pi_N||_{TV} \leq \frac{1}{4}\sum_{j\neq 0} |\hat{K}^n(j)|^2.$$

Then, using the fact that K^n is the n-fold convolution of μ, where $\mu(x) = \frac{1}{3}$ for $x \in \{-1, 0, 1\}$ and zero otherwise, we obtain

$$||K^n - \pi_N||_{TV} \leq \frac{1}{4}\sum_{j\neq 0} \hat{\mu}(j)^{2n}.$$

The transform of μ may be calculated explicitly to be

$$\hat{\mu}(j) = \frac{1}{3} + \frac{2}{3}\cos\left(\frac{2\pi j}{N}\right).$$

There are elementary exponential upper bounds on this from which the upper bound in (23) follows.

For the lower bound, we use an alternative characterisation of the variation distance for two measures, μ and π given by

$$||\mu - \pi||_{TV} = \frac{1}{2}\sup_{||f||_\infty = 1} |\mu(f) - \pi(f)|.$$

Using $f(j) = \cos(2\pi j/N)$ we find that $\pi_N(f) = 0$ and

$$\begin{aligned} K^n(f) &= \sum_j K^n(j)f(j) = \mathbf{Re}\sum_j K^n(j)e_N^j = \mathbf{Re}\hat{K}^n(1) \\ &= \mathbf{Re}\hat{\mu}(1)^n = \left\{\frac{1}{3} + \frac{2}{3}\cos\left(\frac{2\pi}{N}\right)\right\}^n. \end{aligned}$$

Hence, for suitable α, we have

$$||K_N^n(0,\cdot) - \pi_N||_{TV} \geq \frac{1}{2}\left\{\frac{1}{3} + \frac{2}{3}\cos\left(\frac{2\pi}{N}\right)\right\}^n \geq e^{-\frac{\alpha n}{N^2}}.$$

□

Remark: This result gives us that the time to stationarity is of order N^2. Furthermore, using the symmetry of the chain (in particular we have $||K_N^n(0,\cdot) - \pi_N||_{TV} = \max_x ||K_N^n(x,\cdot) - \pi_N||_{TV}$), it may be concluded that there is no variation cutoff.

The case $a = 2$ is more tricky to analyse and is done so for odd N in [3]. The argument used there uses rather fiddly bounds to conclude that the chain reaches stationarity in only $O(\ln N \ln\ln N)$ steps. It is perhaps more illustrative to present an outline of the proof given in [1] that deals with the special case $N = 2^l - 1$ by using a stopping time argument to bound the variation distance.

Let K_N and π_N be the appropriate transition matrix and invariant distribution when $a = 2$.

Theorem

For $N = 2^l - 1$ we have, for $c > 1/\ln 3$,

$$\lim_{l\to\infty} ||K_N^{\lfloor cl\ln l\rfloor}(0,\cdot) - \pi_N||_{TV} = 0.$$

Proof: Aldous/Diaconis prove the result using three lemmas. These are stated here without proof. First though, we define a random variable, $\tilde{U}$, by

$$\tilde{U} = 2^{l-1}\delta_1 + 2^{l-2}\delta_2 + ... + \delta_l$$

where the δ_i are independent, uniform on $\{-1,1\}$. The distribution of $\tilde{U}$ is close to uniform on S_N. Indeed, if we define $P_{\tilde{U}}$ to be the distribution of $\tilde{U}$ it is straightforward to compute that

$$||P_{\tilde{U}} - \pi_N||_{TV} = \frac{1}{2^{l-2}} - \frac{1}{2^l - 1} \leq \frac{1}{2^{l-2}}.$$

We must also introduce a stopping time for the Markov chain defined by (22) with $a = 2$ and $X_0 = 0$. We have

$$X_n = 2^{n-1}b_1 + 2^{n-2}b_2 + ... + b_n (\text{mod } N).$$

Since $N = 2^l - 1$ and $2^l = 1(\text{mod } N)$ this gives

$$X_n = A_1 2^{l-1} + A_2 2^{l-2} + ... + A_l$$

with $A_i = b_i + b_{i+l} + ...$ for $i = 1, ..., l$. T is defined to be the first time that each of the A_i has a non-zero summand. We now state two of the lemmas.

Lemma: The distribution of X_n conditional on $T = m < n$ is the same as the distribution of the $\tilde{U} * V$ where V is a random variable independent of $\tilde{U}$.

Lemma: Let π and μ be two measures on a finite group G. Then

$$||\pi * \mu - |G|^{-1}||_{TV} \leq ||\pi - |G|^{-1}||_{TV}.$$

Now let P^n be the distribution of X_n and P_V be the distribution of V. Applying these two lemmas gives, $\forall m < n$,

$$\begin{aligned} ||P^n(\cdot|T = m) - \pi_N||_{TV} &= ||P_{\tilde{U}} * P_V - \pi_N||_{TV} \\ &\leq ||P_{\tilde{U}} - \pi_N||_{TV} \\ &\leq \frac{1}{2^{l-2}}. \end{aligned} \tag{24}$$

We now state the third lemma which is an adaptation of the bound given at (2).

Lemma: Let $Y_1, Y_2, ...$ be a process taking values in a finite group G. Write Q^n for the distribution of Y_n and S be a stopping time such that, for some $\varepsilon > 0$,

$$||Q^n(\cdot|S = m) - |G|^{-1}||_{TV} \leq \varepsilon, \forall m < n.$$

Then

$$||Q^n - |G|^{-1}||_{TV} \leq \varepsilon + P(S > n).$$

We note that if l is chosen so that $2^{2-l} \leq \varepsilon$, then (24) is precisely the condition of this lemma. Hence, when we revert to the K_N notation, we obtain

$$||K_N^n(0, \cdot) - \pi_N||_{TV} \leq \varepsilon + P(T > n).$$

But bounding the right hand side is fairly straightforward since

$$P(T \overset{\leqslant}{\cancel{>}} ml) \quad = \quad P(\text{At least one of } b_1, b_{1+l}, ..., b_{1+l(m-1)} \neq 0)^l = \left(1 - \left(\frac{1}{3}\right)^m\right)^l$$

We have $c \ln 3 = 1 + \delta$ for some $\delta > 0$. Thus,

$$\begin{aligned} P(T > \lfloor cl \ln l \rfloor) &\leq P(T > l \lfloor c \ln l \rfloor) = 1 - \left(1 - \left(\frac{1}{3}\right)^{\lfloor c \ln l \rfloor}\right)^l \\ &\sim 1 - \left(1 - \left(\frac{1}{3}\right)^{c \ln l}\right)^l = 1 - \left(1 - \frac{1}{l^{\delta+1}}\right)^l \to 0. \end{aligned}$$

From which the claim follows.

□

References

[1] Aldous, D. and Diaconis, P. (1986) *Shuffling cards and stopping times,* Ame. Math. Monthly 93, no. 5, 333-348

[2] Aldous, D. and Diaconis, P. (1987) *Strong uniform times and finite random walks,* Adv. Appl. Math. 8, 69-97

[3] Chung, F., Diaconis, P. and Graham, R. (1987) *Random walks arising in random number generation,* Ann. Appl. Prob. 3, 1148-1165

[4] Diaconis, P. (1996) *The cutoff phenomenon in finite markov chains,* Proc. Natl. Acad. Sci. 93, 1659-1664

[5] Diaconis, P. and Hanlon, P. (1991) *Eigenanalysis for the Metropolis algorithm,* Hypergeometric functions on domains of positivity..., 99-117

[6] Diaconis, P. and Saloff-Coste, L. (1993) *Comparison techniques for random walk on groups,* Ann. Prob. 21, 2131-2156

[7] Diaconis, P. and Stroock, D. (1991) *Geometric bounds for eigenvalues for Markov chains,* Ann. Appl. Prob. 1, 36-61

[8] Fill, J. (1991) *Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains...,* Ann. Appl. Prob. 1, 62-87

[9] Garren, S. and Smith, R. (2000) *Estimating the second largest eigenvalue of a Markov transition matrix,* Bernoulli 6, 215-242

[10] Fulman, J. and Wilmer, E. (1999) *When does Poincaré beat Cheeger?,* Ann. Appl. Prob. 9, 1-13

[11] Norris, J. (1997) *Markov Chains,* Cambridge University Press

[12] Saloff-Coste, L (1997) *Lectures on probability theory and statistics (Saint-Flour, 1996),* Lecture Notes in Math., 1665, Springer, Berlin, 301-413

[13] Sinclair, A. (1992) *Improved bounds for mixing rates of Markov chains and multicommodity flow,* Combinatorics, Probability and Computing, 1, 351-370