http://www.math.nyu.edu/faculty/goodman/teaching/MonteCarlo12/

Class notes: Monte Carlo methods
Week 4, Markov chain Monte Carlo analysis
Jonathan Goodman
February 26, 2013

1 Introduction

Error bars for MCMC are harder than for direct Monte Carlo. It is harder to estimate error bars from MCMC data, and it is harder to predict them from theory. The estimation and theory are more important because MCMC estimation errors can be much larger than you might expect based on the run time.

The fundamental formula for MCMC error bars is as follows. Suppose X_k is a sequence of MCMC samples. We estimate $A = E_f[V(X)]$ using

$$\widehat{A}_n = \frac{1}{n}\sum_{k=1}^{n} V(X_k) \; . \qquad (1)$$

For large n, the variance is

$$\text{var}\Big(\widehat{A}_n\Big) = \frac{\text{var}_f(V(X))}{(n/\tau)} \; . \qquad (2)$$

The new quantity in this formula is the *autocorrelation time*, τ. If $\tau = 1$, this is the error bar formula for direct sampling. In practice, τ can be in the thousands or more, which makes error bars much larger than direct sampling with the same number of samples. This week we will discuss how to estimate τ from Monte Carlo data and how to estimate it in theory.

The autocorrelation time measures the effect of correlations between samples. If $X_0 \sim f$, then $X_k \sim f$ for all $k > 0$. This is what it means for f to be the invariant distribution. But X_k is not independent of X_0 for $k > 0$. We need a quantitative measure of the dependence of X_k on X_0. The *auto-covariance* function is

$$C(n) = \text{cov}_f(V(X_n), V(X_0)) \; . \qquad (3)$$

The subscript f indicates that $X_0 \sim f$. This is the auto-covariance because it is the covariance between the same quantity, $V(X)$, at different times. A *cross covariance* is the covariance of different quantities.

A bad MCMC method has $C(n)$ that converges to zero slowly. Theory identifies (at least) two obstructions to faster convergence: bad geometry and *collective modes*. Bad geometry refers to sets in the state space that are hard to get out of. A simple way to quantify this is

$$P_f(X_{k+1} \notin A \mid X_k \in A) \; .$$

This is small by necessity if $P(A)$ is close to 1. Therefore, we consider the *conductance*

$$\Phi = \inf_{P(A)\leq \frac{1}{2}} P_f(X_{k+1} \notin A \mid X_k \in A) \ . \tag{4}$$

Warning: the term implies that there is a physical motivation to this definition related to electricity or something. To the best of my knowledge, there is no such motivation. The term was invented to sound physical, but it is not.

This definition was motivated by a similar definition given by Jeff Cheeger to quantify "bottlenecks" in manifolds. If A is a set with a significant volume but small boundary, then ∂A (the boundary of A) is a bottleneck. The *Cheeger constant* is

$$C = \inf_{\mathrm{vol}(A)\leq \frac{1}{2}\mathrm{vol}(M)} \frac{\mathrm{vol}_{n-1}(\partial A)}{\mathrm{vol}_n(A)} \ . \tag{5}$$

For 2D ($n = 2$ above), the ratio is the ratio of the arc length of ∂A to the area of A. Cheeger showed that the *spectral gap* of M is at least $\frac{1}{2}C^2$. Similarly, an MCMC method that satisfies detailed balance has a spectral gap at least $\frac{1}{2}\Phi^2$ The definition of conductance (4) will be restated in Section 4 to look more like the Cheeger constant (5).

2 Auto-correlation time, Kubo formula

The *Kubo formula* is

$$\lim_{n\to\infty} n \,\mathrm{var}\Big(\widehat{A}_n\Big) = \sum_{n=-\infty}^{\infty} C(n) \ . \tag{6}$$

Before getting to the not completely rigorous "proof", some remarks. The sum on the right can be thought of as a *diffusion coefficient*

$$D = \sum_{n=-\infty}^{\infty} C(n) \ . \tag{7}$$

The important statement is that for large n,

$$\mathrm{var}\Big(\widehat{A}_n\Big) \approx \frac{D}{n} \ . \tag{8}$$

The *static variance* (more properly, *time zero covariance*) is

$$C(0) = \sigma^2_{V(X)} = \mathrm{var}_f(V(X)) \ .$$

This is a property of the distribution f and the function $V(x)$, not the MCMC method. If we had a direct sampler, then $C(n) = 0$ for $n \neq 0$. This makes the Kubo formula degenerate to

$$\mathrm{var}\Big(\widehat{A}_n\Big) \approx \frac{\mathrm{var}_f(V(X))}{n} \ .$$

This formula is true as an identity, not an approximation.

The covariance function is a symmetric function of n. You can see this from the alternative more constructive definition

$$C(n) = \lim_{k\to\infty} \mathrm{cov}(V(X_{k+n}), V(X_k)) \ . \tag{9}$$

Here we do not assume that $X_0 \sim f$, since it would not be in practice. Section 3 explains why this should be true. Informally, if the distribution of X_k converges to f, and if $n > 0$, then the joint distribution of X_k and X_{k+n} converges to the joint distribution that X_0 and X_n have if $X_0 \sim f$. But the definition (9) makes sense when $n < 0$. And

$$\mathrm{cov}(V(X_{k+n}), V(X_k)) = \mathrm{cov}(V(X_k), V(X_{k+n})),$$

so the limit is the same: $C(n) = C(-n)$.

The *correlation coefficient* is a dimensionless measure of the relationship between two random variables. If V and W are two random variables, then

$$\rho = \mathrm{cor}(V,W) = \frac{\mathrm{cov}(V,W)}{\sqrt{\mathrm{var}\,(V)\mathrm{var}(W)}} \ .$$

This is a number between -1 and 1 (proof: Cauchy Schwarz, applied to $\widetilde{V} = V - E[V]$ and $\widetilde{W} = W - E[W]$). It is the covariance, normalized by V and W in a way that makes ρ dimensionless. The *auto-correlation* function is the un-equal time correlation. The denominator simplifies because $X_0 \sim f$ and $X_n \sim f$, so $\mathrm{var}(V(X_0)) = \mathrm{var}(V(X_n)) = \mathrm{var}_f(V(X))$.

$$\rho(n) = \frac{\mathrm{cov}(V(X_n), V(X_0))}{\mathrm{var}(V(X))} = \frac{C(n)}{C(0)} \ . \tag{10}$$

This gives the alternative definition

$$D = C(0) \sum_{n=-\infty}^{\infty} \rho(n) \ .$$

The *auto-correlation time* is the dimensionless sum (because $\rho(-n) = \rho(n)$)

$$\tau = \sum_{n=-\infty}^{\infty} \rho(n) = 1 + 2\sum_{n=1}^{\infty} \rho(n) \ . \tag{11}$$

We get some insight by rewriting the variance formula for the estimator (8) in terms of the auto-correlation time. The result is (2). The denominator defines the *effective* number of samples

$$n_{\mathit{eff}} = \frac{n}{\tau} \ .$$

The auto-correlation time is the number of MCMC steps that it takes to make one effectively independent sample, for the purpose of estimating A using (1).

Note that τ is a property not only of the MCMC method, but also of the quantity being estimated. It can happen that one quantity is much harder to estimate (has a larger τ) than another. More commonly, the quantity you are interested in has a τ about as large is it can be for the sampler you are using.

In discrete time, the auto-correlation time is dimensionless. But the continuous time version of τ has units of time. If X_t is a continuous time Markov process that has a suitable invariant distribution, then we can look at

$$\widehat{A}_T = \frac{1}{T}\int_0^T V(X_t)\,dt\ .$$

The analogue of the (8) is

$$\mathrm{var}\Big(\widehat{A}_T\Big) \approx \frac{D}{T} = \frac{\mathrm{var}_f(V(X))}{(T/\tau)}\ ,$$

with

$$D = \int_{-\infty}^{\infty} C(t)\,dt\ , \tag{12}$$

and

$$\tau = 2\int_0^{\infty} \rho(t)\,dt\ .$$

Since $\rho(t)$ is dimensionless, τ has units of t. The formula (12) is the more common version of the Kubo formula. It was essentially discovered by Einstein and used in his 1905 paper explaining Brownian motion.

We verify the discrete time Kubo formula (6) under the hypotheses

$$|\mathrm{cov}(V(X_k), V(X_j))| \le Ce^{-\mu|k-j|}\ , \tag{13}$$

and

$$|\mathrm{cov}(V(X_k), V(X_j)) - C(|j-k|)| \le Ce^{-\mu\ \min(j,k)}\ , \tag{14}$$

If course C and μ are positive constants. In Section 3, we verify these hypotheses for discrete state space and non-degenerate MCMC chains. You need more hypotheses to verify it for continuous state space. Normally the covariances are positive, but they do not have to be. There are significant, though rare, practical situation where some covariances are negative. It even is possible to have $\tau < 1$.

Starting with (1), we get

$$\mathrm{var}\Big(\widehat{A}_n\Big) = \frac{1}{n^2}\sum_{j=1}^n\sum_{k=1}^n \mathrm{cov}(V(X_j), V(X_k))\ . \tag{15}$$

We show that

$$n\,\mathrm{var}\Big(\widehat{A}_n\Big) = \frac{1}{n}\left\{\sum_{j=1}^n\left[\sum_{i=-\infty}^{\infty} C(|j-i|)\right] + E\right\}\ , \tag{16}$$

where E is bounded as $n \to \infty$. This will prove the Kubo formula.

There are two differences between the exact sum (15) and the target approximation in (16). One is $\mathrm{cov}(V(X_j), V(X_k))$. The other is expanding the range of the k sum. Both parts reduce to summing a geometric series of error terms. The principle is that a geometric sum is approximately the size of its largest term. More precisely, suppose $x \neq 1$ and $x > 0$, and consider

$$S = \sum_{k=a}^{b} x^k \; .$$

If $x > 1$ or $x < 1$ the largest term is x^b or x^a. Either way, we have

$$S \leq \frac{1}{|1-x|} \max(x^a, x^b) \; .$$

First we bound the error in replacing the non-equilibrium covariances $\mathrm{cov}(V(X_j), V(X_k))$ with the equilibrium values $\mathrm{cov}_f(V(X_j), V(X_k)) = C(|j-k|)$. We seek a bound for

$$E_1 = \sum_{j=1}^{n} \sum_{k=1}^{n} |\mathrm{cov}(V(X_j), V(X_k)) - C(|j-k|)| \; .$$

The summand is symmetric in j and k, so it suffices to bound half the sum

$$E_1 \leq 2 \sum_{j=1}^{n} \sum_{k=1}^{j} |\mathrm{cov}(V(X_j), V(X_k)) - C(|j-k|)| \; .$$

We have assumed two bounds for the summands. One is

$$|\mathrm{cov}(V(X_j), V(X_k)) - C(|j-k|)| \leq Ce^{-\mu \min(j,k)} \; .$$

The other is (with a different C)

$$\begin{aligned} |\mathrm{cov}(V(X_j), V(X_k)) - C(|j-k|)| &\leq |\mathrm{cov}(V(X_j), V(X_k))| + |C(|j-k|)| \\ &\leq Ce^{-\mu|j-k|} \; . \end{aligned}$$

We want to use the better bound, so we see where they are equal, which is

$$\min(j,k) = |j-k|$$

We chose the half sum where $j \geq k$, so this is $k = j - k$, which is $k = j/2$. Therefore:

$$E_1 \leq \sum_{j=1}^{n} \sum_{k=1}^{j/2} e^{-\mu(j-k)} \; + \; \sum_{j=1}^{n} \sum_{k=j/2}^{j} e^{-\mu k} \; .$$

Our general principle for bounding geometric series says that

$$\sum_{k=1}^{j/2} e^{-\mu(j-k)} \leq \frac{1}{|1-e^{\mu}|} e^{-\mu j/2} \; ,$$

and

$$\sum_{k=j/2}^{j} e^{-\mu k} \leq \frac{1}{|1-e^{-\mu}|} e^{-\mu j/2} \ .$$

Therefore

$$E_1 \leq \frac{C}{1-e^{-\mu}} \sum_{j=1}^{n} e^{-\mu j/2} \leq C \ .$$

The constant C is uniform in n as long as μ stays away from zero.

In the second step we expand the range of the k sum. For this, we choose a different inner variable $k = j + i$, $i = k - j$, so that $i = 0$ corresponds to the "diagonal", which is the equal time covariance. The error is

$$\begin{aligned} E_2 &= \left| \sum_{j=1}^{n} \sum_{k=1}^{n} C(|j-k|) - \sum_{j=1}^{n} \sum_{k=-\infty}^{\infty} C(|j-k|) \right| \\ &= \left| \sum_{j=1}^{n} \sum_{i=1-j}^{n-j} C(i) - \sum_{j=1}^{n} \sum_{i=-\infty}^{\infty} C(i) \right| \\ &\leq \sum_{j=1}^{n} \sum_{i=-\infty}^{-j} |C(i)| \ + \sum_{j=1}^{n} \sum_{i=n-j+1}^{\infty} |C(i)| \end{aligned}$$

The two sums on the last line are the same, so we discuss only the first one. That has the bound (recall the largest term principle)

$$\begin{aligned} \sum_{j=1}^{n} \sum_{i=-\infty}^{-j} |C(i)| &\leq C \sum_{j=1}^{n} \sum_{i=-\infty}^{-j} e^{-\mu|i|} \\ &= C \sum_{j=1}^{n} \frac{e^{-\mu j}}{1-e^{-\mu}} \\ &\leq C \frac{1}{(1-e^{-\mu})^2} \ . \end{aligned}$$

This shows that E_2 is bounded as $n \to \infty$. This, and the E_1 bound, prove the Kubo formula (6).

2.1 Estimating D

You can use the Kubo formula to estimate an error bar. You estimate the autocovariance function $C(t)$ then do the sum. But there is a subtlety. If you do too much of the sum, you get an estimator that is *inconsistent.*

Suppose A is a number you want to estimate and $\widehat{A}_n$ is a family of estimators. The family is *consistent* if $\widehat{A}_n$ converges to A as $n \to \infty$ in some sense. Two precise versions of this notion are *weak* consistency and *strong* consistence. The family is weakly consistent if, for any $\varepsilon > 0$,

$$\Pr\Big(\Big|\widehat{A}_n - A\Big| \geq \varepsilon\Big) \;\to\; 0 \;\;, \quad \text{as } n \to \infty \;. \tag{17}$$

This corresponds to $\widehat{A}_n \to A$ *in probability.* As a computational scientist, you would choose the precision, ε and the *confidence* $1-\delta$. Then there is an N_0 so that if $n \geq N_0$, then the estimator error is less than ε with probability at least $1-\delta$ is you take $n \geq N_0$. Moreover, the estimator family is weakly consistent if the bias and variance go to zero as $n \to \infty$. This is easy to check in most practical applications that are weakly consistent.

Strongly consistent is $\widehat{A}_n \to A$ as $n \to \infty$ almost surely. For example, the Kolmogorov law of large numbers states that the sample mean of n i.i.d. samples is a strongly consistent estimator of the population mean. Strong consistency implies weak consistency, but weak consistency does not imply strong consistency. Strong consistency is a nice property to have, but it is harder to prove than weak consistency. As an example, the weak law of large numbers is easier to prove than the strong law.

One way to estimate D is to estimate $C(t)$ and add. Suppose we have an MCMC sequence of samples X_k, and the corresponding time series $Y_k = V(X_k)$, for $k = 1, \ldots, n$. Then

$$\overline{Y}_n = \frac{1}{n}\sum_{k=1}^{n} V(X_k)$$

is the sample mean. The "standard estimator" of the sample auto-covariance is

$$\widehat{C}(t) = \frac{1}{n-t-1}\sum_{k=1}^{n-t}\left(Y_{k+t} - \overline{Y}_n\right)\left(Y_k - \overline{Y}_n\right) \;.$$

We subtract $t+1$ from n in the denominator because there are t terms missing from the sum and because we took away one "degree of freedom" by using $\widehat{Y}_n$ instead of $A = E_f[V(X)]$. In practice, the difference between $n-t-1$ and $n-t$ should not matter at all. The difference between n and $n-t$ barely matters, because t should be much smaller than n.

A natural seeming estimator of D is

$$\widehat{D}_{inc} = \widehat{C}(0) + 2\sum_{t=1}^{n-1}\widehat{C}(t) \;. \tag{18}$$

This estimator is inconsistent because the variance of $\widehat{D}_{inc}$ does not go to zero as $n \to \infty$. You can see this using a straightforward calculation that takes only a few hours. The reason is that although $C(t) \to 0$ as $t \to \infty$, the estimators $\widehat{C}(t)$ all have noise. When you add all the estimators as in (18), you get more noise than signal.

One way to fix this is to cut off the sum.

$$\widehat{D}_W = \widehat{C}(0) + 2\sum_{t=1}^{W} \widehat{C}(t) \; . \tag{19}$$

If the *window size*, W, is too small, $\widehat{D}_W$ is a biased estimator. It is probably biased too low, because the usual case is $C(t) > 0$. This means that your estimated error bars will be too small. If W is too large, $\widehat{D}_W$ is noisy.

The appropriate W depends on the problem. It could be 10 to 10^6. Robust software should estimate an appropriate W from data. One heuristic is the *self consistent window*. Each W gives an estimate of the auto-correlation time by summing the estimated auto-correlation function up to W:

$$\widehat{\tau}_W = \frac{\widehat{D}_W}{\widehat{C}(0)} = 1 + 2\sum_{t=1}^{W} \widehat{\rho}(t) \; . \tag{20}$$

If W is large enough, this is an accurate estimate of the time scale on which $\rho(t)$ decays. A simple strategy is to take the window size to be a multiple of the estimated auto-correlation time

$$W = M\widehat{\tau}_W \; . \tag{21}$$

We typically take $M = 4$ or so. This window is *self consistent* because W is consistent with the τ determined by W. It is found using

```
#define M      5

 double tHat = 1.;
 int    W    = 1;
 while ( tHat < M*tHat )
    tHat += 2*rhoHat[(W++)-1]; // rhoHat[0] is at lag t=1.
```

3 Convergence rate and spectrum

You can learn something about the auto-covariance function using the eigenvalues and eigenvectors of the generator of the MCMC Markov chain. We define the generator first abstractly and then concretely in various specific cases. Spectral theory for abstract operators is complicated, so we discuss some of the possibilities in examples. You do not have to be an expert in functional analysis to do Monte Carlo.

The abstract generator of a discrete Markov chain is a linear operator on the space of functions defined on the state space, $\mathcal{S}$. If $\mathcal{S} = \{1, \ldots, n\}$, then a function defined on $\mathcal{S}$ is a function of $k \in (1, \ldots, n)$, which is a column vector $V \in \mathbb{R}^n$. If $\mathcal{S} = \mathbb{R}^{\backslash}$, then we are talking about functions of n real variables $V(x_1, \ldots, x_n)$. The generator R produces another function $W = RV$ by

$$W(x) = E[V(X_1) \mid X_0 = x] \; . \tag{22}$$

If $\mathcal{S} = \{\infty, \ldots, \backslash\}$ and the transition probabilities are R_{jk}, then $W = RV$ has entries (be careful of our conflicting notation, where x is called j, $W(x)$ is called W_j, etc.)

$$W_j = E[V(X_1) \mid X_0 = j] = \sum_{k \in \mathcal{S}} P(X_1 = k \mid X_0 = j) V_k = \sum_{k=1}^{n} R_{jk} V_k \ . \tag{23}$$

In this case, the abstract generator is the same as the transition matrix. The action of the generator on a vector is matrix multiplication.

Suppose $\mathcal{S} = \mathbb{R}^n$ and we are using a Metropolis sampler with proposal density $Q(x,y)$ and acceptance probability $A(x,y)$. We define $Z(x)$ to be the probability of accepting a proposal made from x. This is

$$Z(x) = \int_y Q(x,y) A(y) \, dy \ .$$

Then $P(X_1 = x \mid X\_0 = x) = 1 - Z(x)$. If $V(x)$ is defined for $x \in \mathbb{R}^n$, and $W = RV$, then

$$\begin{aligned} W(x) &= P(X_1 = x \mid X_0 = x) V(x) + \int Q(x,y) A(x,y) V(y) \, dy \\ W(x) &= (1 - Z(x)) V(x) + \int Q(x,y) A(x,y) V(y) \, dy \ . \end{aligned} \tag{24}$$

You can write this using a transition density $R(x,y)$ if you include a $\delta-$function term to represent the probability of getting a rejection:

$$R(x,y) = \delta(x-y)(1 - Z(x)) + Q(x.y) A(x,y) \ .$$

Then

$$W(x) = \int R(x,y) V(y) \, dy \ .$$

The generator R always has an eigenvector with eigenvalue $\lambda = 1$, the constant function. If $V(x) = 1$ for all x, then (22) gives $W(x) = 1$ for all x. You might wonder about the case $\mathcal{S} = \mathbb{R}^n$, where a constant is not in L^2. We will see that $V \equiv 1$ is in L^2 when viewed the right way.

Any linear operator has a set of complex numbers associated to it, the *spectrum*. If R is an $n \times n$ matrix, the spectrum is the set of all eigenvalues of R. If R is an integral operator, odds are R is compact. The spectrum of a compact operator consists of all its eigenvalues and the number $\lambda = 0$, even if 0 is not an eigenvalue. Many operators, including (24) are not compact. For those, the spectrum is the complement of the *resolvent set*. A complex number λ is in the resolvent set if $(R - \lambda I)^{-1}$ is a well defined bounded operator. This means that there is a positive C so that for any W, there is a unique V with $\|V\| < \infty$, and so that $RV - \lambda V = W$. This V satisfies[1] $\|W\| \leq C \|V\|$.

[1] The resolvent set, and the spectrum, may depend on the norm $\|V\|$. There is a natural norm that is relevant for the auto-covariance function, as we will see.

Spectral theory for linear operators is more complicated that eigenvalue-eigenvector theory for square matrices because there is more to spectrum than eigenvalues. Eigenvalues are in the spectrum, as defined in the previous paragraph. If there is a $V \neq 0$ with $\|V\| < \infty$ so that $RV - \lambda V = 0$, then the solution of $RV - \lambda V = 0$ is not unique. If there is a family of eigenvalues $\lambda_j \to \lambda$ as $j \to \infty$, then λ is in the spectrum of R even if it is not a proper eigenvalue. You can see this by looking at the solution of $RU_j - \lambda U_j = V_j$. The solution is

$$U_j = \frac{1}{\lambda - \lambda_j} V_j \ .$$

This means that there is no finite C so that $\|U_j\| \leq C \, \|V_j\|$. A compact operator typically ("typically" → "always" if $\|V\|$ is a Hilbert space norm) has eigenvalues $\lambda_j \to 0$ as $j \to \infty$. For the Metropolis sampler operator (24), all the numbers $Z(x)$ are in the spectrum and it is likely that most of them are not eigenvalues.

With luck, our MCMC operator R has a positive *spectral gap*, which is

$$\mu = \min\{1 - |\lambda| \ \text{ with } \lambda \in \operatorname{spec}(R) \ , \ \ \lambda \neq 1\} \ . \tag{25}$$

The spectral gap is the distance in the complex plane from the largest non-trivial eigenvalue to the unit circle. The smaller the spectral gap, the slower $C(t)$ can converge to 0, and the larger the auto-correlation time can be.

Let us come back to the concrete case of an n state Markov chain with transition matrix R. We denote the solution of the eigenvalue problem as

$$Rr_j = \lambda_j r_j \quad , \quad l_j R = \lambda_j l_j \quad , \quad l_j r_k = \delta_{jk} \ .$$

The r_j and l_j are right and left eigenvectors respectively. The last relation is *bi-orthogonality*. By convention $\lambda_1 = 1$ and $|\lambda_j| < 1$ if $j \neq 1$. We assume there are no non-trivial Jordan blocks. If there are Jordan blocks the formulas are more complicated but the conclusions are about the same. By convention, $l_1 = f$ (the steady state probability distribution) and $r_j = \mathbf{1}$ (the vector of all ones). The relation $l_1 r_1 = 1$ is $\sum_j (f_j \cdot 1) = 1$.

Let V an *observable*, some function of $\mathcal{S}$ we are interested in. The expected value is

$$\overline{V} = E_f[V] = \sum_{j=1}^{n} f_j V_j \ .$$

The static variance is

$$C(0) = \operatorname{var}(V) = \sum_{j=1}^{n} \left(V_j - \overline{V}\right)^2 \ .$$

The lag t covariance is calculated using the lag t transition probabilities (You can prove this by induction on t starting from the lag 1 case that is the definition of R.)

$$P(X_t = k \mid X_0 = j) = R^t_{jk} \ .$$

To calculate the result, use the joint probability formula

$$P_f(X_t = k \text{ and } X_0 - j) = P(X_t = k \mid X_0 = j)P(X_0 = j) = R^t_{jk} f_j \ .$$

$$\begin{aligned}
C(t) &= \text{cov}_f(V(X_t), V(X_{}))) \\
&= E_f[(V(X_t) - \overline{V})(V(X_0) - \overline{V}) \\
&= \sum_{j=1}^{n} \sum_{k=1}^{n} P_f(X_t = k \text{ and } X_0 - j) \left(V_k - \overline{V}\right) \left(V_j - \overline{V}\right) \\
C(t) &= \sum_{j=1}^{n} \sum_{k=1}^{n} f_j R^t_{jk} \left(V_k - \overline{V}\right) \left(V_j - \overline{V}\right) \ . \qquad (26)
\end{aligned}$$

We express this formula in a more natural way. A natural inner product for functions on $\mathcal{S}$ is

$$\langle W, V \rangle_f = E_f[W(x)V(x)] = \sum_{j=1}^{n} f_j W_j V_j \ . \qquad (27)$$

In this notation, the auto-covariance function is

$$C(t) = \langle \left(V - \overline{V}\mathbf{1}\right) , \, R^t \left(V - \overline{V}\mathbf{1}\right) \rangle_f \ . \qquad (28)$$

You find the behavior for large t using the right eigenvector basis. If W is any observable, then there are coefficients a_j so that

$$W = \sum_{j=1}^{n} a_j r_j \ .$$

The coefficients are given by

$$a_j = l_j W \ .$$

Since $l_1 = f$,

$$a_1 = l_1 W = fW = \sum_{j=1}^{n} f_j W_j = E_f[W] \ .$$

Of course,

$$R^t W = \sum_{j=1}^{n} \lambda_j^t a_j r_j \ .$$

But $|\lambda_j| < 1$ except $\lambda_1 = 1$, so $R^t W \to E_f[W]$ as $t \to \infty$.

The formula for $C(t)$ in (28) involves $R^t\left(V - \overline{V}\mathbf{1}\right)$. The expected value of $V - \overline{V}\mathbf{1}$ is zero (check: $fV = \overline{V}$), so

$$V - \overline{V}\mathbf{1} = \sum_{j=2}^{n} a_j r_j \ .$$

We find that

$$C(t) = \sum_{j=2}^{n} b_j \lambda_j^t \ , \tag{29}$$

where

$$b_j = \langle V - \overline{V}\mathbf{1}, r_j \rangle_f \ . \tag{30}$$

Since $|\lambda_j| \leq 1 - \mu$ for $j \geq 2$ (μ being the spectral gap), we see that the sum $D = \sum C(t)$ converges and the auto-correlation time is finite.

You can get a feel for auto-correlation time by supposing λ_2 (say) is extremal among decaying modes, so $|\lambda_2| = 1 - \mu$. Suppose $b_2 = 1$ and $b_j = 0$ otherwise. Then $C(0) = 1$, so (11) gives the auto-correlation time as

$$\tau = \sum_{-\infty}^{\infty} (1-\mu)^t = \left[2 \sum_{t=0}^{\infty} (1-\mu)^t \right] - 1 = 2\frac{1}{\mu} - 1 = \frac{2-\mu}{\mu} \approx \frac{2}{\mu} \ .$$

A small spectral gap leads to a large auto-correlation time and slow MCMC convergence.

The auto-correlation time depends on both the MCMC Markov chain and the observable. The analysis of the previous paragraph is not sharp, see Exercise 1. Eigenvalue/eigenvector analysis is not always a good way to understand high powers of large matrices.

The analysis is easier when the MCMC chain satisfies detailed balance. That is because detailed balance implies that R is self adjoint in the f inner product (27). In general, a matrix A is self adjoint with respect to inner product $\langle \cdot, \cdot \rangle$ if

$$\langle AW, V \rangle = \langle W, AV \rangle$$

for every pair of vectors W and V. For example, A is self adjoint in the l^2 inner product,

$$\langle W, V \rangle_{l^2} = \sum_{j=1}^{n} W_j V_j \ ,$$

if and only if A is symmetric: $A_{jk} = A_{kj}$ for all j and k. The detailed balance condition $P_f(\text{observe } j \rightarrow k) = P_f(\text{observe } k \rightarrow j)$ is

$$f_j R_{jk} = f_k R_{kj} \ . \tag{31}$$

Then, R is self adjoint in the f inner product if

$$\langle RW, V \rangle_f = \langle W, RV \rangle_f \ .$$

This is easy to write out

$$\begin{aligned} \sum_{j=1}^{n} (RW)_j \, f_j V_j &= \sum_{k=1}^{n} W_k f_k \, (RV)_k \\ \sum_{j=1}^{n} \sum_{k=1}^{n} R_{jk} W_k f_j V_j &= \sum_{k=1}^{n} \sum_{j=1}^{n} W_k f_k R_{kj} V_j \ . \end{aligned}$$

The detailed balance relation (31) makes this last line true.

4 Conductance and spectral gap

5 Examples and exercises

1. Find an example of a non-degenerate Markov chain with spectral gap μ and an observable V so that the corresponding auto-correlation time satisfies $\tau_V \gg 2\mu^{-1}$.