Lecture Notes on Monte Carlo Methods
Fall Semester, 2005
Courant Institute of Mathematical Sciences, NYU
Jonathan Goodman, goodman@cims.nyu.edu

Chapter 5: Dynamic sampling and Markov chain Monte Carlo.

A variety of techniques collectively called[1] *Markov chain Monte Carlo* (MCMC) or *dynamic sampling* allow sampling of complex high dimensional distributions not accessable by simple samplers. With technical ideas to follow, the rough idea is that if f is the invariant law of a nondegenerate Markov chain $X(t)$, then the law of $X(t)$ converges to f as $t \to \infty$. We choose a starting state $X(0)$ in an arbitrary way and *run* the chain. Regardless of the distribution of $X(0)$, the law of $X(t)$ will converge to f as $t \to \infty$. Moreover, for large enough s, the samples $X(t)$ and $X(t+s)$ "decorrelate" (become independent, actually) so as to become effectively distinct samples of f. Therefore we can estimate $A = E_f[V(X)]$ using

$$\widehat{A} = \frac{1}{L} \sum_{t=1}^{L} V(X(t)) \; . \tag{1}$$

Regardless of the choice of $X(0)$, we have $\widehat{A} \to A$ as $L \to \infty$. This is called the *ergodic theorem* for Markov chains, in anology to ergodic theorems in statistical mechanics. It is remarkable that there are many distributions that are impractical to sample directly but easy (or at least possible) to sample with MCMC.

A rejection sampler is inefficient if its acceptence probability is low. A dynamic sampler can be inefficient because of large long lasting correlations between $X(t)$. The *autocorrelation time* is the minimum s before $X(t)$ and $X(t+s)$ are effectively independent samples from the point of view of error bars. A *rapidly mixing* chain has small autocorrelation time. There are tricks for inventing rapidly mixing chains in various situations (e.g. Multigrid Monte Carlo, many more in Alan Sokal's notes or Jun Liu's book). We will discuss error bars and autocorrelation time estimates for dynamic Monte Carlo.

All of the simple sampling tricks apply to dynamic MCMC sampling, but there are three more: *detailed balance*, *partial resampling* (also called the *Gibbs sampler*[2] and *composition*. A Markov chain with transition matrix P samples f if the balance equatinos $fP = f$ are satisfied. Designing a MCMC sampler means seeking stochastic *moves*, or random transitions $X(t) \to X(t+0)$, that preserve f, which means that if $X(t) \sim f$ then $X(t+1) \sim f$. The Metropolis and partial resampling methods are simple tricks for finding such random moves. A Markov chain must be *ergodic* if the law of $X(t)$ is to converge to f from any

[1] The practice predates the name by several decades.

[2] The term Gibbs Sampler is used by statisticians, either because it was used before them to sample from the Gibbs Boltzmann distribution in systems such as the Ising model, or because in such sampling problems, you need a simple sampler for a simple conditional Gibbs distribution.

starting distribution. This means roughly that there is a nonzero probability path (sequence of states) connecting any two allowed states: any allowed state is reachable from any other allowed state. We often need to combine or compose several different moves to get a Markov chain that is ergodic as well as preserving f. Moreover, using a richer family of elementary moves can give a Markov chain with faster mixing.

1 Finite State Space

Suppose $\mathcal{S}$ is a set with d elements, x_1, ..., x_d. The number d usually will be enormous, exponential in the size of the problem. For example, we might have n *spin variables* that take the values $+1$ or -1. The set of all possible spin states would be $d = 2^n$.

We want to sample a distribution determined by the probabilities $f(x)$. These are organized into a length d row vector, $f = (f_1, \ldots, f_d)$ with $f(k) = f(x_k) = P(X = x_k$. A stationary Markov chain is characterized by the *transition probabilities* $P_{xy} = P(x \to y)$ $P(X(t+1) = x \mid X(t) = y)$. These transition probabilities may be organized into a $d \times d$ matrix P with entries $P_{jk} = P_{x_j, x_k}$. We often call the set of probabilities a *transition rules*. The transition rules preserve f, or leave f *invariant*, if

$$\{\ P(X(t) = x) = f(x) \ \text{ for all } x \in \mathcal{S}\ \} \implies \{\ P(X(t+1) = x) = f(x) \ \text{ for all } x \in \mathcal{S}\ \}\ .$$

This is expressed as a *balance condition*

$$f(x) = \sum_{y \in \mathcal{S}} f(y) P_{yx} \quad \text{for all } x. \tag{2}$$

In matrix terms, this is[3]

$$f_k = \sum_j f_j P_{jk}\ , \quad f\ fP\ . \tag{3}$$

Outside of Monte Carlo, it is common to have the transition rules P and use them to find the invariant probabilities f. In MCMC the situation is reversed. We have f and seek P to satisfy the balance conditions (2). Although f is uniquely determined by P (if P is nondegenerate), there are many different sets of transition rules, P, compatible with a given f. Paradoxically, this can make it harder to find a suitable P. It helps to seek restricted classes of moves or more restrictive balance conditions.

1.1 Detailed balance

The transition probabilities P_{xy} satisfy *detailed balance* if in the steady state the probability of observing an $x \to y$ transition is equal to the probability of

[3]This equation makes sense with f before P because f is a row vector.

observing $y \to x$. To observe $x \to y$, the system first must be in state x, then it must choose the $x \to y$ transition. The probability of this is $f(x)P_{xy}$. Therefore, the detailed balance condition is

$$f(x)P_{xy} = f(y)P_{yx} \quad , \quad \text{for all } x, y. \tag{4}$$

The balance condition states that each state probability is in balance with all the others collectively. In particular, a system that satisfies detailed balance with a given f also satisfies the balance condition (2) with the same f. Indeed, if we sum over y in (4) and use the fact that

$$\sum_y P_{xy} = 1 \quad , \quad \text{for all } x,$$

(because P is a transition matrix of a Markov chain), we get (2). Thus, the detailed balance condition is one recipe for balance. If we are given probabilities $f(x)$ and are able to find transition rules that satisfy detailed balance, then these transition rules also satisfy the ordinary balance conditions (2) and leave f invariant. The resulting Markov chain will be ergodic if the transition rules are powerful enough.

1.2 The Metropolis method

The Metropolis method[4] is a way to find probabilities P_{xy} that satisfy detailed balance. An example illustrates the reasoning. Let $\mathcal{S} = \{1, \ldots, d\}$ and let the probabilities $f(x)$ be given. Try to find a *random walk*[5] with $a(x) = P(x \to x+1)$, $b(x) = P(x \to x)$, and $c(x) = P(x \to x-1)$. The constraints are

$$a(x) + b(x) + c(x) = 1 \quad \text{for all } x, \tag{5}$$

and

$$0 \le a(x) \le 1 \quad , \quad 0 \le b(x) \le 1 \quad , \quad 0 \le c(x) \le 1 \quad \text{for all } x. \tag{6}$$

We disallow transitions that would take us outside of $\mathcal{S}$: $P(d \to d+1) = a(d) = 0$, $P(1 \to 0) = c(1) = 0$. We require $a(x) > 0$ and $c(x) > 0$ otherwise, for otherwise the Markov chain would have noncommunicating components.

The detailed balance condition (4) for this case is

$$f(x)a(x) = f(x+1)c(x+1) \; . \tag{7}$$

This is one equation for the two unknowns $a(x)$ and $c(x+1)$. In order to have a rapidly mixing chain, we try to choose $a(x)$ and $c(x)$ as close to one as possible consistent with the constraints (5) and (6).

[4]The method was proposed in a paper by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (two pairs of relatives) and sometimes is called the $\mathrm{MR}^2\mathrm{T}^2$ method.

[5]Random walks with non constant transition probabilities may be called *birth and death* processes. If so, $X(t)$ represents the nomber present at time t and $x \to x+1$ is a birth and $x+1 \to x$ is a death.

- case (i), $f(x) \le f(x+1)$. Rewrite (7) as

$$c(x+1) = \frac{f(x)}{f(x+1)} a(x) \ , \tag{8}$$

which will be satisfied if

$$(wrong) \qquad a(x) = 1 \ , \quad c(x+1) = \frac{f(x)}{f(x+1)} \ . \qquad (wrong) \tag{9}$$

- case (ii), $f(x) \ge f(x+1)$. Rewrite (7) as

$$a(x) = \frac{f(x+1)}{f(x)} c(x) \ , \tag{10}$$

which will be satisfied if

$$(wrong) \qquad c(x+1) = 1 \ , \quad a(x) = \frac{f(x+1)}{f(x)} \ . \qquad (wrong) \tag{11}$$

These formulas would determine all the $a(x)$ and $c(x)$ and we use (5) to get $b(x)$. The trouble is that we probably find $a(x) + c(x) > 1$ for some x, which gives $b(x) < 0$, which is not a probability.

One cure for this is to be less *greedy*[6]. If we make sure that $a(x) \le \frac{1}{2}$ and $c(x) \le \frac{1}{2}$, then (5) never will give $b(x) < 0$. The revised method is:

- If $f(x) \ge f(x+1)$, take

$$c(x+1) = \frac{1}{2} \ , \quad a(x) = \frac{f(x+1)}{f(x)} \frac{1}{2} \tag{12}$$

- If $f(x) \le f(x+1)$, take

$$a(x) = \frac{1}{2} \ , \quad c(x+1) = \frac{f(x)}{f(x+1)} \frac{1}{2} \tag{13}$$

- Take $a(d) = 0$, $c(1) = 0$.
- Take $b(x) = 1 - a(x) - c(x)$.

These rules define a Markov chain that satisfies detailed balance for the probabilities $f(x)$.

We reinterpret this to uncover the idea behind the Metropolis method. The formula (12) may be viewed as saying first propose $x \to x+1$ with probability $\frac{1}{2}$, then accept this move with probability $\frac{f(x+1)}{f(x)} \frac{1}{2}$. If the move is rejected than $X(t+1) = x$. Since $f(x) > f(x+1)$, a chain with f as a steady state should

[6]Greedy is a technical term in computer science for algorithms that make choices that seem optimal at one state without regard to how these choices effect future choices. Jeff Rauch pointed out that *short sighted* is more apt, but this cannot reverse a decades old habit.

spend more time at x than at $x+1$. The rejection step achieves this by rejecting sometimes the $x \to x+1$ transitions and staying at x instead. On the other hand, if $f(x) > f(x+1)$, then a proposed $x+1 \to x$ step never is rejected.

In general, a Metropolis sampler has a *proposal* and a *rejection* stage. The proposal move from x to y has probability K_{xy} with $K_{xy} \geq 0$ for all x, y and $\sum_y K_{xy} = 1$. The acceptance probabilities R_{xy} are chosen to incure detailed balance for f. The probability of observing an $x \to y$ ($x \neq y$) transition is $f(x)K_{xy}R_{xy}$ (be at x, propose $x \to y$, accept $x \to y$). The detailed balance consition (4) is

$$f(x)K_{xy}R_{xy} = f(y)K_{yx}R_{yx} \; . \tag{14}$$

As before, this is one equation for the two unknowns R_{xy} and R_{yx}, the constrainst being $0 \leq R_{xy} \leq 1$ and $0 \leq R_{yx} \leq 1$. Again we make the transition probabilities as large as possible by choosing one of them to be 1 and the other determined by (14). This leads to the Metropolis formula for acceptance probabilities

$$R_{xy} = \min\left(1, \frac{f(y)K_{yx}}{f(x)K_{xy}}\right) \; . \tag{15}$$

The proposal probabilities K_{xy} are arbitrary and the acceptance probabilities are adjusted using (15) to give detailed balance. In practice, we choose the K_{xy} to be simple and easy to implement while trying to make the chain mix reasonably well.

1.3 Metropolis for the Ising model

The Ising model is a simple microscopic model of a magnetic material. The hypothesis is that electrons orbiting the atoms in a crystal (say, of iron) form little magnets the either point up (*spin* $= +1$) or down (*spin* $= -1$). If a majority of these microscopic spins point in the same direction, the crystal has a net macroscopic magnetic field.

Let Λ be a two dimensional square *lattice* with *sites* (j,k) for $j = 1, \ldots, n$ and $k = 1, \ldots, n$. A *spin configuration*, x, is a function of (j,k) with $x_{jk} = \pm 1$, assigning a spin to each site $(j,k) \in \Lambda$. Site (j,k) is a *nearest neighbor* of site (j',k') if either $j' = j+1$, $k' = k$, or $j' = j$, $k' = k+1$. A nearest neighbor pair forms a *bond*. This bond is *aligned* if $x_{jk} = x_{j'k'}$ and *misaligned* otherwise. The *energy* of x is the number of misaligned bonds, which may be written

$$H(x) = \sum_{\langle (j,k),(j',k') \rangle} \frac{1}{2}\left(1 - x_{jk}x_{j'k'}\right) \; . \tag{16}$$

The notation $\langle (j,k),(j',k') \rangle$ means that (j,k) and (j',k') are nearest neighbors. The summand equals zero if the spins are aligned and one if they are not. The Gibbs Boltzmann probabilities are

$$f(x) = \frac{1}{Z(\beta)} e^{-\beta H(x)} \; , \tag{17}$$

where the *partition function* is

$$Z(\beta) = \sum_{x \in \mathcal{S}} e^{-\beta H(x)} \ .$$

The parameter β is the *inverse temperature*[7].

A simple Metropolis strategy for the Ising model proposes to *flip*[8] a single randomly chosen spin, with the proposal being accepted or rejected according to detailed balance. This proposal rule satisfies its own balance relation

$$K_{xy} = K_{yx} \ . \tag{18}$$

If x and y differ by a spin then $K_{xy} = K_{yx} = 1/n^2$ (n^2 being the number of spins, each of which is equally likely to be chosen). Otherwise (two or more spin differences), $K_{xy} = K_{yx} = 0$. In this case, the acceptance formula (15) simplifies to[9]

$$R_{xy} = \min\left(1, \frac{f(y)}{f(x)}\right) \ . \tag{19}$$

When we use the Gibbs Boltzman probabilities (17) in (19) or (15), the common partition function factor $Z(\beta)$ drops out. This is just as well since we usually do not know it. In fact, many properties of physical systems modeled by Gibbs Boltzmann probabilities are determined by the partition function and its derivatives. If we know the partition function we would not need Monte Carlo. In general, the Metropolis method depends only on ratios of probabilities rather than the probabilities themselves. In practice, the reason for using symmetric proposal probabilities (18) is that if $K_{xy} \neq K_{yx}$, then they probably themselves have normalization factors that are hard to know.

The Metropolis strategy can be described directly in terms of the physical energy function $H(x)$. Starting from a given configuration, x, we propose modified configuration, y. We calculate the old and new energies, $H(x)$ and $H(y)$, and the change in energy, $\Delta H = H(y) - H(x)$. If $\Delta H < 0$, then y is more likely than x and we accept y. If $\Delta H > 0$, the old configuration is more likely, so we accept y with probability $e^{-\beta \Delta H} < 1$ and otherwise stay at x.

This Metropolis strategy for the Ising model gives an ergodic Markov chain. In fact, two configurations x and y can differ in at most n^2 places so n^2 spin flip steps can take x to y with positive probability. However, in some applications it will be hard for the practitioner to distinguish between 2^{-100} and exact mathematical zero.

---

[7] In terms of physical temperature, T, $f(x) = Z^{-1} e^{H(x)/kT}$, where k is the Boltzmann constant which has units of energy per degree. If we were using physical units, there would be a dimensional parameter in (16) also.

[8] If (j,k) is the chosen lattice site, we flip its spin by $y_{jk} = -x_{jk}$. The remaining spins are unchanged, $y_{j'k'} = x_{j'k'}$ if $(j,k) \neq (j',k')$.

[9] The more complicated formula (15) sometimes is called the *Metropolis Hastings* formula. The $\mathrm{MR^2T^2}$ paper only had the special case (19).

1.4 Partial resampling and heat bath

The Ising model spin configuration is a collection of individual spin variables x_{jk}. *Partial resampling*, also called the *heat bath*[10] method, in its simplest form, replaces x_{jk} by an independent sample $\widetilde{x}_{jk}$ with the correct conditional distribution. More precisely, write $x = (x_{jk}, x')$, where

$$x' = \{x_{j'k'} \mid (j',k') \neq (j,k)\} \ ,$$

and take $\widetilde{x} = (\widetilde{x}_{jk}, x')$ – leave all spins unchanged except possibly for x_{jk}.

It seems clear that partial resampling preserves the equilibrium distribution f. Going through the formal argument, if $x \sim f$ then $x' \sim f'_{jk}$, the marginal distribution of x' (this being the definition of marginal distribution). Bayes' rule states that if x' has the correct marginal distribution and if $\widetilde{x}_{jk}$ has the correct conditional distribution (conditioned on x', then $(\widetilde{x}_{jk}, x')$ has the correct overall distribution. This is exactly what partial resampling accomplishes.

The single site heat bath partial resampling method is easy to implement for the Ising model. Let $x_\pm = (\pm 1, x')$, be the spin configurations with $x_{jk} = \pm 1$, and let $h_\pm = H(x_\pm)$ be the corresponding energies, wtih $\Delta H = h_+ - h_-$. Resampling x_{jk} means sampling from the two possibilities $\widetilde{x}_{jk} = \pm 1$ with probabilities $f_\pm = Cf(x_\pm)$. We know neither $f(x_+)$ not $f(x_-)$, but their ratio is

$$\frac{f_+}{f_-} = \frac{f(x_+)}{f(x_-)} = e^{-\beta \Delta H} \ .$$

Therefore

$$\begin{cases} f_+ = P\left(x_+ \mid x'\right) = \dfrac{e^{-\beta\Delta H}}{1 + e^{-\beta \Delta H}} \\ \\ f_1 = P\left(x_+ \mid x'\right) = \dfrac{1}{1 + e^{-\beta \Delta H}} \ . \end{cases} \tag{20}$$

this is implemented by

```
if ( rand() < 1/(1+e^{-βΔH}) ) x_jk= -1;

else x_jk = +1;
```

The heat bath probability of a spin flip is higher than the Metropolis probability (reader: check this). Thus, the heat bath method is slightly better than Metropolis and has the same complexity. One would use Metropolic in situations where partial resampling is impractical.

Partial resampling expresses a general strategy for sampling of complex high dimensional distributions – reduce them to a sequence of low dimensional dimensional sampling problems accessible to simple samplers. This is anologous

[10]The name comes from the physical picture that we resample x_{jk} by touching the (j,k) lattice site to a *heat bath*. Physicists imagine contact with a heat bath as a way to bring a system into its equilibrium Gibbs Boltzmann distrigution.

to solving large systems of equations by iterative methods, but with an importance difference. If we want to solve $F(x) = b$, we can use any heuristics we like, provided that at the end of the day we evaluate the residual to check whether $F(x) = b$. Unfortunately, there seems to be no Monte Carlo anologue of the residual. Therefore, whetever moves we use (Metropolis, partial resampling, ..), we must design them to preserve f.

2 Error analysis for dynamic Monte Carlo

The error analysis for the dynamic Monte Carlo estimator (1) must take into account correlation between samples. for large enough run length, this amounts to replacing L by the *effective sample size*

$$L_{\mathit{eff}} = L/\tau \ , \tag{21}$$

where τ is the *correlation time*[11]. In turn, τ is given by the *Kubo formula*[12] in terms of the *autocorrelation function*, $\rho(s)$ (defined and explained below):

$$\tau = \sum_{s=-\infty}^{\infty} \rho(s) = 1 + 2\sum_{s=1}^{\infty} \rho(s) \ . \tag{22}$$

For a fixed s, the equilibrium *autocovariance* at lag s is

$$C(s) = \lim_{t\to\infty} \operatorname{cov}\left(V(X(t)), V(X(t+s))\right) \ . \tag{23}$$

As $t \to \infty$, the distribution of $X(t)$ converges to f, so for $s > 0$,

$$C(s) = \operatorname{cov}_f\left(V(X(0)), V(X(s))\right) \ ,$$

where cov_f means that we choose $X(0) \sim f$, something we can do in theory but not in practice. The correlation coefficient, ρ, is a dimensionless version of covariance:

$$\operatorname{cor}(U, W) = \frac{\operatorname{cov}(U, W)}{\sqrt{\operatorname{var}(U) \cdot \operatorname{var}(W)}} \ .$$

Take $U = V(X(t))$ and $W = V(X(t+s))$ and note that if $X(t) \sim f$ then $\operatorname{var}(X(t)) = \operatorname{var}(X(t+s))$, so the lag s correlation is

$$\rho(s) = \lim_{t\to\infty} \operatorname{cor}\left(V(X(t)), V(X(t+s))\right) = \lim_{t\to\infty} \frac{\operatorname{cov}\left(V(X(t)), V(X(t+s))\right)}{\operatorname{var}\left(V(X(t))\right)} \ . \tag{24}$$

Again, for $s > 0$, this is the same as

$$\rho(s) = \operatorname{cor}_f\left(V(X(0)), V(X(s))\right) = \frac{\operatorname{cov}_f\left(V(X(0)), V(X(s))\right)}{\operatorname{var}_f\left(V(X)\right)} \ .$$

[11] This is related to but distinct from measures of equilibration time, the time needed for the distribution of $X(t)$ to reach approximate steady state.

[12] This should be called the *Einstein Kubo* formula because the main idea is in Einstein's 1905 paper explaining Brownian motion that won him the Nobel Prize.

Clearly $\rho(0) = 1$ and (from (23) $C(-s) = C(s)$ and $\rho(-s) = \rho(s)$.

The significance of τ is its role in the error bar of the estimator (1) (using $\sigma(\cdot)$ to represent the standard deviation):

$$\sigma(\widehat{A}_L) \approx \frac{1}{\sqrt{L\tau}}\sigma_f(V(X)) \; . \tag{25}$$

In the case of independent samples, the Kubo formula (22) gives $\tau = 1$. Computations with $\tau > 100$ (not uncommon) have error bars at least an order of magnitude larger than incorrect independent sample error bars. What matters is not the run length, but the effective sample size (21).

The calculation behind (25) is (using $t' = t + s$, $s = t' - t$):

$$\begin{aligned}
\text{var}\left(\widehat{A}_L\right) &= \text{var}\left(\frac{1}{L}\sum_{t=1}^{L} V(X(t))\right) \\
&= \frac{1}{L^2}\sum_{t=1}^{L}\sum_{t'=1}^{L} \text{cov}\left(V(X(t)), V(X(t'))\right) \\
&= \frac{1}{L^2}\sum_{t=1}^{L}\sum_{s=1-t}^{L-t} \text{cov}\left(V(X(t)), V(X(t+s))\right) \; .
\end{aligned}$$

In this sum we suppose (see *Decay of correlations*, below) that for large L, the vast majority of terms have t and $t + s$ so large that we may replace $\text{cov}\left(V(X(t)), V(X(t+s))\right)$ by its limit $C(s)$. We also suoppose that for most t, the difference between

$$\sum_{s=1-t}^{L-t} C(s) \quad \text{and} \quad \sum_{-\infty}^{\infty} C(s)$$

is negligable. Also, using (22),

$$\sum_{-\infty}^{\infty} C(s) = \text{var}_f(V(X)) \sum_{-\infty}^{\infty} \frac{C(s)}{C(0)} = \tau \, \text{var}_f(V(X)) \; .$$

With all this,

$$\begin{aligned}
\text{var}\left(\widehat{A}_L\right) &\approx \sum_{t=1}^{L} \tau \, \text{var}_f(V(X)) \\
&= \frac{\tau}{L}\text{var}_f(V(X)) \; ,
\end{aligned}$$

which is (25).

3 Decay of correlations

In some cases we can understand the decay of correlations using eigenvalues and eigenvectors. First, the joint distribution of $(X(0), X(s))$ is given (in equilibrium) by

$$\begin{aligned} P_r\left(X(0) = x, X(s) = y\right) &= P_f\left(X(0) = x\right) \cdot P\left(X(s) = y \mid X(0) = x\right) \\ &= f(x) P_{xy}^s \ . \end{aligned}$$

Next, assuming (as we may) that $E_f\left[V(X)\right] = 0$,

$$\begin{aligned} C(s) &= E_f\left[V(X(0))V(X(s))\right] \\ &= \sum_{xy} V(x)V(y)f(x)P_{xy}^s \ . \end{aligned} \tag{26}$$

In matrix terms, this is

$$C(s) = (Vf)^t \, P^s V \ , \tag{27}$$

where Vf is the d component column vector with entries $V(x)f(x)$. If we suppose P is diagonalizable[13] and write

$$Pr_j = \lambda_j r_j \quad , \quad j = 1, \ldots, d \ ,$$

where r_j are linearly independent right eigenvectors. Since P is the transition matrix for a Markov chain, $\sum_y P_{xy} = 1$ for all x and therefore the eigenvector $r_1 = (1, 1, \ldots, 1)^t$ has eigenvalue $\lambda_1 = 1$. Part of the Perron Frobenius theorem states that $|\lambda_j| < 1$ for $j \geq 2$ if the Markov chain is nondegenerate.

Let R be the $d \times d$ matrix whose columns are the right eigenvectors, starting with r_1. The inverse, $L = r^{-1}$ has rows that are left eigenvectorsL $l_j P = \lambda_j l_j$. The eigenvector corresponding to $\lambda_1 = 1$ is $l_1 = f$, the invariant distribution. $(f(y) = \sum_x f(x) P_{xy})$. Note that $l_1 \cdot r_1 = \sum_x f(x) \cdot 1 = 1$. We write V in terms of right eigenvectors

$$V = \sum_{j=2}^{d} a_j r_j \ . \tag{28}$$

The term corresponding to r_1 is missing because its coefficient is zero. In fact, $a_j = l_j \cdot V$, and in particular,

$$a_1 = l_1 \cdot V = \sum_x f(x) V(x) = E_f\left[V(X)\right] = 0 \ .$$

Putting this back into (27) gives

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

[13] Tf P had nontrivial Jordan blocks, the discussion is more wordy but otherwise about the same.

where

$$b_j = a_j \, (Vf)^t \, r_j \ .$$

A first interpretation of the *spectral representation* (29) makes use of the *spectral gap*

$$\alpha = 1 - \max_{j>1} |\lambda_j| > 0 \ .$$

Since $|\lambda_j| \leq 1 - \alpha$ for $j \neq 1$, we have[14]

$$|C(s)| \leq C \, (1-\alpha)^{|s|} \ . \tag{30}$$

Thus for a finite state space nondegenerate Markov chain, the autocovariance function decause exponentially with s. If P has Jordan blocks then (30) may change to

$$|C(s)| \leq C \left(1 + |s|^k\right) (1-\alpha)^{|s|} \ ,$$

which still implies exponential decay.

The spectral gap and decay bound (30) may be used to define another heuristic measure of the mixing rate of a Markov chain. If we neglect the constant in (30) and sum,

$$\frac{1}{\alpha} = \sum_{s=0}^{\infty} (1-\alpha)^s \ ,$$

we get the *exponential*, or *spectral* decorrelation time $\tau_{\exp} = 1/\alpha$. The units are clearer in continuous time, where the Kubo formula (22) would be

$$\tau = \int_{-\infty}^{\infty} \rho(s) ds \ ,$$

and (30) would be $C(s) \leq C \exp(-s/\tau_{\exp}$. By anology with the integral formula, the Kubo autocorrelation time may be called the *integrated* autocorrelation time. In continuous time, both the integrated and the spectral times have units of time. People often say that the exponential time based on the spectral gap is important because it controls the rate of convergence of the distribution of $X(t)$ to f. I am skeptical.

It may be naive to rely on (30) in practical situations because the constants may grwo with d (which may be very large). In fact, the graph of $C(s)$ as a function of s may look nothing like a decaying exponential. See work on *threshhold phenomena* by Persi Diaconis and collaborators or *non normal matrices* by Nick Trefethen and collaborators. Practical Markov chains used in Monte Carlo, particularly those with detailed balance, tend not to be as bad as the worst counterexamples.

[14] This uses $C(-s) = C(s)$ to extend to negative s.

4 Detailed balance

The spectral analysis that led to (30) is more powerful for Markov chains that satisfy detailed balance because detailed balance implied that the transition matrix, P, is similar to a symmetric matrix. The detailed balance condition (4) multiplied by $f(x)^{1/2} f(y)^{-1/2}$ gives

$$\widetilde{P}_{xy} = f(x)^{1/2} P_{xy} f(y)^{-1/2} = f(y)^{1/2} P_{yx} f(x)^{-1/2} = \widetilde{P}_{yx} \; . \tag{31}$$

The covariance formula (26) becomes

$$\begin{aligned} C(s) &= \sum_{xy} \widetilde{V}(x) \widetilde{P}_{xy} \widetilde{V}(y) \\ &= \widetilde{V}^t \widetilde{P} \widetilde{V} \; , \end{aligned} \tag{32}$$

where $\widetilde{V}(x) = f(x)^{1/2} V(x)$. Because $\widetilde{P}$ is symmetric we know that all the eigenvalues λ_j are real (though possibly negative), that there are no nontrivial Jordan blocks, and that the right eigenvectors of $\widetilde{P}$ may be chosen to be orghonormal:

$$\widetilde{P} \widetilde{r}_j = \lambda_j \widetilde{r}_j \quad , \quad \widetilde{r}_j^t \widetilde{r}_k = \delta_{jk} \; .$$

If we define an f inner product

$$\langle u, w \rangle_f = \sum_x f(x) u(x) w(x) \; ,$$

then the right eigenvectors of P are orthonoemal in the f inner product:

$$\langle r_j, r_k \rangle_f = \sum_x f(x) r_j(x) r_k(x) = \widetilde{r}_j^t \widetilde{r}_k = \delta_{jk} \; .$$

Therefore we may write $V = \sum_j a_v r_j$ with $a_j = \langle r_j, V \rangle_f$ (or equivalently, $\widetilde{V} = \sum_j a_j \widetilde{r}_j$ where $a_j = \widetilde{r}_j^t \widetilde{V}$). Either way (32) gives

$$C(s) = \sum_{j=2}^{d} a_j^2 \lambda_j^{|s|} \; . \tag{33}$$

Moreover, the orthonormality of eigenvectors implies that

$$\sum_{j=2}^{d} a_j^2 = \widetilde{V}^t \widetilde{V} = \sum_x f(x) V(x)^2(x) = \mathrm{var}_f(V(X)) \; . \tag{34}$$

All this implies that

- $C(2s) \geq 0$.

(autocovariances and autocorrelations of even offset are positive.)

- $|C(2s+k)| \le C(2s)$ for $k > 0$.

(monotone decay of correlations from even values)

- $|C(s)| \le \mathrm{var}_f(V(X))\,(1-\alpha)^{|s|}$

(an explicit control of the constant that was hard to pin down in (30))

- $|\rho(s)| \le (1-\alpha)^{|s|}$

Sum this to get a rigorous bound that does not require forgetting a constant:

$$\tau \le \frac{2}{\alpha} \,. \tag{35}$$

The spectral gap gives a specific bound for the autocorrelation time that does not depend on dimension. Moreover the bound (35) does not depend on the function $V(x)$, it applies to any estimation problem with a given Markov chain. For this reason, most theoretical analysis of performance of Markov chains with detailed balance works by estimating the spectral gap.

5 Estimating τ

The problem of estimating τ is subtle and I do not know a good solution. To understand the difficulties, let us proceed naively. Given a run of length L, we could compute the sample mean (1) then estimate the lag s covariance by

$$\widehat{C}(s) = \frac{1}{L-s} \sum_{t=1}^{L-s} \left(V(X(t)) - \widehat{A} \right) \left(V(X(t+s)) - \widehat{A} \right) \,, \tag{36}$$

and $\widehat{\rho}(s) = \widehat{C}(s)/\widehat{C}(0)$. So far, so good. The trouble comes with, say,

$$(\textit{wrong}) \qquad \widehat{\tau} = 1 + 2 \sum_{t=1}^{L} \frac{\widehat{C}(s)}{\widehat{C}(0)} \,. \qquad (\textit{wrong}) \tag{37}$$

Statisticians say an estimator is *consistent* if if it converges to the right answer as $L \to \infty$. In particular, the estimator (37) is *strongly* consistent if $\widehat{\tau}_L \to tau$ as $L \to \infty$ (almost surely). *Weak* consistency means that, for any $\epsilon > 0$, $P(|\widehat{\tau}_L - tau| \ge \epsilon) \to 0$ as $L \to \infty$. Part of the Borel Cantelli lemma implies that an estimator that is not weakly consistent cannot be strongly consistent, almost sure convergence implies convergence in probability. We argue that the estimator (37) is not weakly consistent.

In practice, one usually studies weak consistency or inconsistency using the mean and variance. The estimator is *asymptotically unbiased* if $E\left[\widehat{\tau}_L\right] \to E\left[\tau\right]$ as $L \to \infty$. Some pencil scratching shows that our (37) is asymptotically unbiased. If $\widehat{\tau}_L$ is asymptotically unbiased, it will be weakly consistent if $\mathrm{var}(\widehat{\tau}_L) \to 0$ as $L \to \infty$ (by Chebychev's inequality). On the contrary, we argue that $\mathrm{var}(\widehat{\tau}_L)$

does not converge to zero as $L \to \infty$. Let us neglect the uncertainty in $\widehat{C}(0)$. From (**??**), if the $X(t)$ were independent the variance of $\widehat{C}(s)$ is of order $1/L$. Generally speaking, correlated $X(t)$ would make matters worse. The sum (37) has L terms each with variance of order $1/L$. If the terms were independent (they are not), the sum would have variance of order one. With correlated $\widehat{C}(s)$, the variance is at least as big. Finally, with large sample size, the central limit theorem implies that the $\widehat{C}(s)$ are approximately normal (though correlated), so $\widehat{\tau}_L$ also is asymptotically normal. As an asymptotically normal sequence of random variables, $\widehat{\tau}_L$, must have $\text{var}(\widehat{\tau}_L) \to 0$ as $L \to \infty$ in order to converge in probability. This makes it pretty clear that $\widehat{\tau}_L$ is not a consistent estimator, even in the weak sense.

What goes wrong is that most of the terms in (37) have more noise than signal. The expected value of $\widehat{C}(s)$ decays to zero exponentially (30) but the variance does not go to zero as $L \to \infty$. If we *truncate* the sum (37):

$$\widehat{\tau}_L = 1 + 2 \sum_{s=1}^{M} \frac{\widehat{C}(s)}{\widehat{C}(0)} \,, \tag{38}$$

we may be able to adjust M to get most of the signal and not too much noise. If M is too small, we have a biased estimate (nearly always biased toward unrealistically small error bars). If M is too large, the noise swamps the signal. The problem of estimating

$$D = \sum_{-\infty}^{\infty} C(s) \tag{39}$$

(a close relative of τ) is a special case of a central problem in statistics about which several books and countless articles have been written, spectral density estimation. Nevertheless, there does not seem to be a good way to do it.

A good estimator of τ or D should not ask the user to presupply the answer. It also should work over a wide range of τ values. On one extreme, we may have a Markov chain with detailed balance and a spectral gap $\alpha = .5$. Then we know (see the reasoning for (35)) that $M = 10$ should get most of the signal (relative bias less than $1/2^{10}$). On the other extreme are problems with long tails (slowly decaying signal) and $\tau > 1000$. In my work I generally use a *self consistent window* (the truncated sum (38) is a *window*). The goal is to take $M = 5\tau$. If the signal decays exponentially, this introduces a relative bias e^{-5}. Since we don't know τ in advance, we take the smallest instead take $M = 5\widehat{\tau}$. The resulting estimate is *self consistent* (some would say circular) because it requires the estimate of τ to be consistent with the value of τ used in the estimation. In practice, I start with a small trial M and keep increasing M as until (38) gives a $\widehat{\tau}$ with $M \geq 5\widehat{\tau}$. I believe (but never have proven) that this estimator is consistent (in the limit $5 \to infty$). More importantly, it seems to be robust. for error bars, it is more important to be reliably within, say, 20% than to be highly accurate in the possibly impractical limit $L \to \infty$.

6 Composition

A Monte Carlo *move* is a transition matrix that preserves f. We may have several strategies that lead to different moves, say $P^{(1)}$ and $P^{(2)}$ (not to be confused with powers of P). In this case, applying first $P^{(1)}$ then $P^{(2)}$ is a valid composite move because it too preserves f (twice). The corresponding transition matrix is $P^{(1)}P^{(2)}$. Any number of moves may be composed and in any order.

One class of examples is single site resamplers for the Ising model. Let (j,k) be a specific lattice site, and let $P^{(j,k)}$ represent heat bach resampling of the spin at (j,k)[15] One way to possibly flip each spin is to choose a site at random. This has transition matrix

$$P_{\mathrm{RS}} = \frac{1}{n^2} \sum_{(j,k)} P^{(j,k)} \; . \tag{40}$$

Since each of the $P^{(j,k)}$ satisfies detailed balance, so does P_{RS}. The RS is for *random scan*, which may be an self contradiction. It also is possible to visit the sites in a definite order, for example *lexicographically*: $(1,1)$, $(1,2)$, ..., $(1,n)$, $(2,1)$, ..., (n,n). The transition matrix for this is

$$P_L = P^{(1,1)} P^{(1,2)} \cdots P^{(1,n)} P^{(2,1)} \cdots P^{(n,n)} \; .$$

It is natural to compare the effectiveness of n^2 spin flips of random sites

$$P_{\mathrm{RS}}^{n^2}$$

to n^2 spin flips in one systematic scan of the lattice. In the case of a "massless free field", it is possible to do the analysis and see that systematic scan is about twice as good as random scan.

[15]The reader should verify that there is no point in resampling the same site twice – $(P^{(j,k)})^2 = P^{(j,k)}$. Also, if $Q^{(j,k)}$ represents a Metropolis trial at site (j,k), then $(Q^{(j,k)})^r \to P^{(j,k)}$ as $r \to \infty$.