Lecture 19: Markov Chain Monte Carlo

36-350, Fall 2011

2 November 2011

1 Mixing and Correlation Time

Let's suppose for simplicity that our Markov chain has only one eigenvector v with eigenvalue 1, so all the other eigenvalues are strictly < 1 in magnitude. As we saw last time, if we put the eigenvalues and their eigenvectors in decreasing order,

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_K| \tag{1}$$

then when we evolve some initial distribution p_0 for many time steps, the second largest eigenvalue dominates:

$$\begin{aligned} p_0 \mathbf{q}^t &= \sum_{i=1}^{K} a_i v_i \lambda_i^t & (2)\\ &\approx v + a_2 v_2 \lambda_2^t & (3) \end{aligned}$$

or

$$p_0 \mathbf{q}^t - v \approx a_2 v_2 \lambda_2^t \tag{4}$$

So the magnitude of the distance from equilibrium is

$$||p_0 \mathbf{q}^t - v|| \approx |a_2||\lambda_2|^t \tag{5}$$

(since all the eigenvectors have norm 1).

How many steps τ does it take for the Markov chain to come close, say within a distance h, of the equilibrium distribution? Let's appeal to Eq. 5:

$$\begin{aligned} h &= ||p_0 \mathbf{q}^\tau - v|| & (6)\\ &\approx |a_2||\lambda_2|^\tau & (7)\\ \log h &\approx \log|a_2| + \tau \log|\lambda_2| & (8)\\ \tau &\approx \frac{\log h - \log|a_2|}{\log|\lambda_2|} & (9) \end{aligned}$$

Notice that since h and $|\lambda_2|$ are both < 1, their logarithms are negative numbers, and this τ ends up positive, as it should.

τ is called the **mixing time**, because it indicates how long we have to wait for the chain to mix together different initial conditions, and forget which state it started

in. This matters because, while the approach to equilibrium is exponentially fast, the base of the exponent is $|\lambda_2|$, and if that is too close to 1, we will see very little movement[1].

1.1 Correlations

The mixing time can be somewhat pessimistic if we are interested only in a certain function of the state, and not the whole distribution of states. Let's fix our function f and look at the random sequence $Y_1 = f(X_1), Y_2 = f(X_2), \ldots$. In general, this sequence will not be a Markov chain, but we can still say something about its convergence.

Make three assumptions:

1. The expectation value is constant:
$$\mathbb{E}[Y_1] = \mathbb{E}\left[Y_t\right] = \mu \tag{10}$$
2. Covariances are "stationary":
$$\mathrm{Cov}\left[Y_t, Y_s\right] = \rho(|t-s|) \tag{11}$$
3. Covariances are "summable":
$$\sum_{h=0}^{\infty} |\rho(h)| = \varkappa\rho(0) < \infty \tag{12}$$

We are interested in the time average
$$\frac{1}{n}\sum_{t=1}^{n} Y_t \tag{13}$$
The expectation of this is constant:
$$\begin{aligned} \mathbb{E}\left[\frac{1}{n}\sum_{t=1}^{n} Y_t\right] &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[Y_t\right] &(14)\\ &= \mu &(15)\end{aligned}$$
The variance is a little trickier:
$$\begin{aligned} \mathrm{Var}\left[\frac{1}{n}\sum_{t=1}^{n} Y_t\right] &= \frac{1}{n^2}\mathrm{Var}\left[\sum_{t=1}^{n} Y_t\right] &(16)\\ &= \frac{1}{n^2}\left[\sum_{t=1}^{n}\mathrm{Var}\left[Y_t\right] + 2\sum_{t=1}^{n-1}\sum_{s=t+1}^{n}\mathrm{Cov}\left[Y_t, Y_s\right]\right] &(17)\\ &= \frac{\rho(0)}{n} + \frac{2}{n^2}\sum_{t=1}^{n-1}\sum_{s=t+1}^{n}\rho(s-t) &(18)\\ &= \frac{\rho(0)}{n} + \frac{2}{n^2}\sum_{t=1}^{n-1}\sum_{h=1}^{n-t}\rho(h) &(19)\end{aligned}$$

[1] Remember that for small x, $(1-x)^n \approx 1 - nx$, so if $\lambda_2 = 1 - 10^{-12}$, say, then $\lambda_2^{1000} \approx 1 - 10^{-9}$.

Now we use some inequalities:

$$\begin{aligned}
\sum_{h=1}^{n-t} \rho(h) &\leq \sum_{h=1}^{n-t} |\rho(h)| && (20)\\
&\leq \sum_{h=1}^{\infty} |\rho(h)| && (21)\\
&= \varkappa \rho(0) && (22)
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{n}\sum_{t=1}^{n} Y_t\right] &\leq \frac{\rho(0)}{n} + \frac{2n}{n^2}\varkappa\rho(0) && (23)\\
&= \frac{\rho(0)}{n}\left[1+2\varkappa\right] && (24)\\
&= \frac{\rho(0)}{n/\tau_c} && (25)
\end{aligned}$$

where $\tau_c = 1 + 2\varkappa$ is the **correlation time**. With uncorrelated samples, $\tau_c = 1$ and we get the usual behavior we're used to. With correlated samples, we get much the same sort of behavior, but the effective number of samples is not n but n/τ_c.

Since $\frac{\rho(0)}{n/\tau_c} \to 0$ as $n \to \infty$, the variance of the time-average shrinks, and so it converges on the true expectation μ.

Problem

Suppose that $\mathbb{E}\left[Y_t\right]$ is not constant, but that $\mathbb{E}\left[Y_t\right] \to \mu$ as $t \to \infty$. What has to change in the argument above? What if it is only

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[Y_t\right] \to \mu \, ? \tag{26}$$

Problem

Suppose that $Y_{t+1} = \phi Y_t + \epsilon_t$, where $|\phi| < 1$ and the ϵ_t are independent variables with expectation 0 and variance σ^2. Find the $\rho(h)$ function, and calculate $\varkappa$.

2 Convergence of Continuous Markov Processes

We have analyzed finite-state Markov chains because the math needed to understand them is fairly elementary. The math needed to give a parallel analysis of Markov processes with infinitely many states, or with continuous states, is significantly harder. One of the main issues is that a finite Markov chain must eventually hit a recurrent component and stay there, but that doesn't have to happen in infinite state spaces.

(There may be no recurrent states, for one thing.) Still, let's say a little bit about continuous-state Markov processes.[2]

First of all, instead of having a transition matrix $\mathbf{q}$, we have a transition density $q(y|x)$, which gives the conditional probability density of X_{t+1} at y, given that $X_t = x$. To evolve a density, we use an integral rather than matrix multiplication:

$$p_{t+1}(x) = \int p_t(x')q(x|x')dx' \tag{27}$$

A function f is an **eigenfunction**, with eigenvalue λ, when

$$\int f(x')q(x|x')dx' = \lambda f(x) \tag{28}$$

The eigenvalues of a transition density q are all still inside the unit circle. Invariant probability densities p^* are the eigenfunctions with eigenvalue 1,

$$p^*(x) = \int p^*(x')q(x|x')dx \tag{29}$$

but they may not exist.

Roughly speaking, for there to be an invariant distribution, we need to find a part of the state space with three properties:

1. The region is itself invariant: once the process enters, it never leaves.
2. No strictly smaller region is also invariant.
3. After some fixed and finite number of steps, there is a positive probability of going from any part of the region to any other[3].

When these conditions hold, then the ergodic theorem holds, because any trajectory of the Markov process ends up wandering over the whole invariant region, and spending the same amount of time in each part of it as any other trajectory.

As for the rate of convergence of time averages, the arguments in Section 1.1, about the convergence of any one function in terms of its correlation time, go through just as before. Replicating the analysis of mixing times is a bit trickier. Remember that with a finite number K of states, there are K eigenvectors. As $K \to \infty$, the number of eigenvectors, and of eigenvalues, therefore goes towards infinity. Continuous Markov processes therefore usually have infinitely many eigenvalues and eigenfunctions. It can still happen that there is a gap between 1 and the next largest eigenvalue; in that case, everything happens more or less as before, and we have exponentially-fast convergence to the equilibrium distribution. Unfortunately, it can happen that there are infinitely many eigenvalues arbitrarily close to 1, and then convergence to the equilibrium is less than exponential.

[2]For an accessible over-view of continuous Markov processes, see [1]. For the gory details, see [6].

[3]In symbols, there is an k such that, for any sets A, B in the region, $\mathbb{P}\left(X_{t+k} \in B | X_t \in A\right) > 0$. Notice that we need to have *one* k which works for *all* pairs of sub-regions.

3 Markov Chain Monte Carlo

The ergodic theorem tells us that when $X_1, X_2, \ldots X_t, \ldots$ come from a reasonable Markov process,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} f(X_i) = \mathbb{E}_v[f(X)] \tag{30}$$

with v being the invariant distribution of the Markov process. We can read this equation in either direction: if it is easy for us to calculate expectations under v, we "read it from the right", and it gives us a short-cut for long simulations. If on the other hand it is easier for us to simulate, we "read it from the left", and it gives us a way to do complicated integrals.

One very important case of "reading from the right" is the way search-engines calculate the "page-rank" of web documents, the amount of time a random walk on the Web would spend on a given page. This is a bit long to go into here, but see `http://www.stat.cmu.edu/~cshalizi/350/lectures/03/03.pdf`.

For "reading from the left" to be useful, we would need to come up with a Markov process whose invariant distribution was our target distribution v, without being able to just draw independent samples from v. This might sound odd, but it is actually very common. This is because a lot of distributions have the form

$$p(x) \propto f(x) \tag{31}$$

for some nice function f. This is fine, but the integral of f is usually not 1. So we need

$$p(x) = \frac{f(x)}{\int_y f(y)dy} \tag{32}$$

We would need to find the integral in the denominator to actually sample from p, but it's generally not very easy to compute. For instance, you remember that

$$\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi} \tag{33}$$

but (if you are anything like most students!) do not remember why. You will find it instructive to try to find

$$\int_0^{\infty} x^{-\alpha} e^{-\lambda x} \tag{34}$$

as a function of α and λ — and these are comparatively easy cases.

To give a somewhat concrete illustration, think of error-correction in communications. Your friend wants to send you a message, x. What you receive is a distorted and noisy version of it, z. You can try to recover the original by asking how probable different messages were. By Bayes's rule,

$$p(x|z) = \frac{p(z|x)p(x)}{p(z)} = \frac{p(z|x)p(x)}{\int_y p(z|y)p(y)dy} \tag{35}$$

The numerator here is fairly tractable: $p(x)$ represents how common different signals are, and $p(z|x)$ represents the noise and distortion. But the denominator is an ugly thing, where you would have to integrate over all possible signals — and then do that *again* as soon as z changed.

It would be nice if there was some way we could sample from distributions like this without needing to know the normalizing factors (the integrals in the denominator). This is where "Markov chain Monte Carlo", a.k.a. "the Metropolis algorithm" [5], comes in.

Suppose that p is our favorite probability density, which we want to sample from, and $p(x) = f(x)/c$, where the constant c is one of the awkward things we want to avoid calculating. If we could find a Markov process with transition density q, where

$$p(x)q(y|x) = p(y)q(x|y) \tag{36}$$

then it is not hard to convince yourself that p would be invariant. This is because the probability flowing out of state x to state y is exactly balanced by probability flowing into x from y, so nothing changes. (A distribution p obeying Eq. 36 is say to "obey detailed balance".) Now let's try to solve this for q:

$$\frac{q(y|x)}{q(x|y)} = \frac{p(y)}{p(x)} = \frac{f(y)}{f(x)} \tag{37}$$

and the awkward normalizing constant has disappeared.

This leads to **the Metropolis algorithm**:

1. Set X_1 however we like, and initialize $t \leftarrow 1$.
2. Draw a **proposal** Z_t from some conditional distribution $r(\cdot|X_t)$ — for instance a Gaussian or uniform centered on X_t, or anything else easy to draw from and with smooth enough noise.
3. Draw U_t independently from a uniform distribution on $[0, 1)$.
4. If $U_t < f(Z_t)/f(X_t)$, then $X_{t+1} = Z_t$, otherwise $X_{t+1} = X_t$.
5. Increase t by 1 and go to step 2.

What we return as the sample is the sequence of X_t values.

You will notice that the Metropolis algorithm is very like the rejection method for generating random variables, but not quite. In particular, if $f(Z_t) > f(X_t)$, then $X_{t+1} = Z_t$ automatically. The chain always accepts proposals which move it towards regions of higher density. It *sometimes* accepts proposals which move it towards lower density — it's very likely to accept them if $f(Z_t)$ is only a little below $f(X_t)$, and never accepts them if $f(Z_t) = 0$.

You can check that

- The X_t variables are a Markov process, since the distribution of X_{t+1} depends only on the value of X_t.

- The transition density obeys Eq. 37.
- Consequently, Eq. 36 holds, and the invariant distribution of the Markov process is p, as desired.

Since p is invariant, if we drew X_1 according to p, we would always be sampling from the desired distribution. Of course we don't know how to do that, so we draw X_1 from some other, more convenient distribution. This means that the first samples are from the wrong distribution, and we shouldn't use them. (We need to let the Markov chain "burn in".) How many values should we discard from the beginning of the X sequence? That depends on how rapidly the chain mixes, which is why we went over that earlier. There are a lot of diagnostic calculations which are supposed to tell whether the chain has adequately mixed. The simplest of these (though computationally expensive) is to run multiple copies of the chain independently, and see what the time averages have converged.

Problem

It is common to model the time between events like industrial accidents, earthquakes, or e-mails with exponential distributions[4]. That is, the probability that the time between two events falls between x and $x + dx$ is $\approx \lambda e^{-\lambda x} dx$. Unfortunately, the rate λ could itself change. One way to model this is to have λ be random. Suppose that λ follows a gamma distribution with scale a and shape 1. Then

$$p(x) = \int_0^\infty d\lambda \frac{\lambda^{a-1} e^{-\lambda}}{\Gamma(a)} \lambda e^{-\lambda x} \tag{38}$$

It would be handy to infer the current value of the rate λ from the observed times between events. Set up a Markov chain to sample from the distribution of rates conditional on the most recent inter-event time x,

$$p(\lambda|x) = \frac{p(x|\lambda)p(\lambda)}{p(x)} \tag{39}$$

but do not solve Eq. 38 for $p(x)$. (Assume that a is known.) Then check your by solving the equation and getting the exact formula for $p(\lambda|x)$, and comparing that density to the histogram from your code for several values of x. What would you have to change in your code to use a different, non-gamma prior distribution of rates λ?

Hint: The variable-rates model, the exact calculation with a gamma distribution of rates, and comparison to accident data all come from [2]. For application of closely related models to e-mail, without the gamma distribution, see [4, 3].

[4] Empirically, this is because it often works; theoretically, it is because a continuous-time Markov process spends an exponentially-distributed amount of time in each state before moving.

References

[1] Grimmett, G. R. and D. R. Stirzaker (1992). *Probability and Random Processes*. Oxford: Oxford University Press, 2nd edn.

[2] Maguire, B. A., E. S. Pearson and A. H. A. Wynn (1952). "The Time Intervals between Industrial Accidents." *Biometrika*, **39**: 168–180. URL `http://www.jstor.org/pss/2332475`.

[3] Malmgren, R. Dean, Jake M. Hofman, Luis A. N. Amaral and Duncan J. Watts (2009). "Characterizing Individual Communication Patterns." In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*. URL `http://arxiv.org/abs/0905.0106`.

[4] Malmgren, R. Dean, Daniel B. Stouffer, Adilson E. Motter and Luís A. N. Amaral (2008). "A Poissonian explanation for heavy tails in e-mail communication." *Proceedings of the National Academy of Sciences (USA)*, **105**: 18153–18158. URL `http://arxiv.org/abs/0901.0585`. doi:10.1073/pnas.0800332105.

[5] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller (1953). "Equations of State Calculations by Fast Computing Machines." *Journal of Chemical Physics*, **21**: 1087–1092.

[6] Meyn, S. P. and R. L. Tweedie (1993). *Markov Chains and Stochastic Stability*. Berlin: Springer-Verlag.