The Metropolis-Hastings Algorithm

1 Markov Chain Notation for a Continuous State Space

A sequence of random variables $X_0, X_1, X_2, \ldots$, is a **Markov chain on a continuous state space** if...

... where it goes depends on where is is but not where is was.

I'd really just prefer that you have the "flavor" here. The previous Markov property equation that we had

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j | X_n = i)$$

is uninteresting now since both of those probabilities are zero when the state space is continuous.

It would be better to say that, for any $A \subseteq \mathcal{S}$, where S is the state space, we have

$$P(X_{n+1} \in A | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} \in A | X_n = i).$$

Note that it is okay to have equalities on the right side of the conditional line. Once these continuous random variables have been observed, they are fixed and nailed down to discrete values.

1.1 Transition Densities

The continuous state analog of the one-step transition probability p_{ij} is the **one-step transition density**. We will denote this as

$$p(x, y).$$

This is not the probability that the chain makes a move from state x to state y. Instead, it is a probability density function in y which describes a curve under which area represents probability. x can be thought of as a parameter of this density.

For example, given a Markov chain is currently in state x, the next value y might be drawn from a normal distribution centered at x. In this case, we would have the transition density

$$p(x, y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-x)^2}.$$

The analogue of the n-step transition probability $p_{ij}^{(n)}$ is the **n-step transition density** denoted by $p^{(n)}(x, y)$.

Note that we must have

$$\int_{\mathcal{S}} p(x, y)\, dy = 1 \qquad \text{and} \qquad \int_{\mathcal{S}} p^{(n)}(x, y)\, dy = 1$$

for all $n \geq 1$.

1.2 Chapman-Kolmogorov Stuff

Suppose that the chain is currently at state x. The location of the chain 2 time steps later could be described by conditioning on the intermediate state visited at time 1 as follows.

$$p^{(2)}(x,y) = \int_{\mathcal{S}} p(x,z) \cdot p(z,y)\, dz.$$

Note the independence of transitions. Once you are at the intermediate state z, your move to y is independent of your previous move from x to y.

In general, for any $0 \leq m \leq n$, we have

$$p^{(n)}(x,y) = \int_{\mathcal{S}} p^{(m)}(x,z) \cdot p^{(n-m)}(z,y)\, dz.$$

(We define $p^{(0)}(x,y)$ as the Dirac delta function.)

1.3 The Stationary Distribution

Let $\{X_n\}_{n\geq 0}$ be a Markov chain living on a continuous state space $\mathcal{S}$ with transition probability density $p(x,y)$.

Definition: A stationary distribution for $\{X_n\}$ on $\mathcal{S}$ is a probability density function $\pi(x)$ on $\mathcal{S}$ satisfying

$$\pi(y) = \int_{\mathcal{S}} \pi(x)\, p(x,y)\, dx.$$

Note that this is a direct analogue to the discrete state space stationary equation. It again describes the situation

$$X_0 \sim \pi \qquad \Rightarrow \qquad X_1 \sim \pi \qquad \Rightarrow \qquad X_2 \sim \pi \qquad \Rightarrow \qquad \cdots$$

for continuous random variables X_0, X_1, etc... Again, the transitive nature of this relation implies that we also have

$$\pi(y) = \int_{\mathcal{S}} \pi(x)\, p^{(n)}(x,y)\, dx$$

for any fixed n.

1.4 A Limiting Distribution

Can you possibly imaging what this section is about?

Given a continuous state Markov chain with n-step transition densities $p^{(n)}(x,y)$, suppose that the following limit exists and is independent of x.

$$\lim_{n\to\infty} p^{(n)}(x,y).$$

Then the limit is a probability density function in y that is also a stationary distribution. (The proof is very similar to the discrete state-space case of the earlier handout.)

"Well behaved" Markov chains have such a limiting distribution and also are guaranteed to have a unique stationary distribution. In this case, they are one and the same!

1.5 Reversibility (Detailed Balance)

Definition: A Markov chain on a continuous state space $\mathcal{S}$ with transition probability density $p(x, y)$ is said to be **reversible** with respect to a density $\pi(x)$ if

$$\pi(x)\, p(x, y) = \pi(y)\, p(y, x) \tag{1}$$

for all $x, y \in S$. This is also referred to as a **detailed balance** condition.

While it is not required that a Markov chain be reversible with respect to its stationary distribution, it is true that any distribution π satisfying (1) is, in fact, stationary since

$$\pi(x)\, p(x, y) = \pi(y)\, p(y, x)$$

$$\Downarrow$$

$$\int_{\mathcal{S}} \pi(x)\, p(x, y)\, dx = \int_{\mathcal{S}} \pi(y)\, p(y, x)\, dx = \pi(y) \underbrace{\int_{\mathcal{S}} p(y, x)\, dx}_{1} = \pi(y). \ \surd$$

2 The Metropolis-Hastings Algorithm

In this Bayesian course, we are going to need to sample/simulate values from various posterior distributions. We have talked about a few methods of simulating from distributions already but each had some requirement that may or not be true of our distribution of interest. There is no "one size fits all" algorithm for simulating random variables.

Markov chain Monte Carlo (MCMC) is a large class of algorithms that one might turn to where one creates a Markov chain that converges, in the limit, to a distribution of interest. For example, if one wanted to draw/simulate values from a particular posterior density $\pi(\theta|\vec{x})$ (note the totally optional switch to a more Markov looking notation for this density), an MCMC algorithm might give you a recipe for a transition density $p(\cdot, \cdot)$ that walks around on the support of $\pi(\theta|\vec{x})$ so that

$$\lim_{n\to\infty} p^{(n)}(\cdot, \theta) = \pi(\theta|\vec{x}).$$

The Metropolis-Hastings algorithm is one such algorithm.

2.1 The Setup

We will call any density we want to simulate values from a "target density". Suppose that we have a target density $\pi(x)$. Here, x may be discrete or continuous, and also may be high-dimensional. In what follows, I will only use continuous state space notation.

In order to run the Metropolis-Hastings algorithm to simulate values from $\pi(x)$, we need to choose "any" transition density $q(x, y)$. This "arbitrary" choice often has nothing to do with π. The point of the Metropolis-Hastings algorithm is to make an accept-reject type

adjustment so that we are ultimately running a simulation of a Markov chain with some transition density $p(x, y)$ that is reversible with respect to the target π. This chain will then have π as a stationary distribution and will be well behaved enough to only have one and to have a limiting distribution. Thus, they will be one and the same so we will iterate our simulation forward in time in an event to reach the limiting/stationary distribution.

If one is clever, one might not make the choice of $q(x, y)$ completely arbitrary, but instead will choose something for which the chain will converge faster to π. Often, one is not able to be clever. :(

Common categories for $q(x, y)$ to fall in are

- symmetric: $q(x, y) = q(y, x)$ (This is the pure Metropolis algorithm.)
- independent: $q(x, y) = q(y)$ (Your proposal is independent of where you are.)

but neither of these are necessary.

2.2 The Algorithm

To get approximate draws from π, start a chain at an arbitrary value for X_0. Let $n = 0$. Suppose that $X_n = x$.

1. Propose a value y drawn from $q(x, y)$.
2. With probability
$$\alpha(x, y) := \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right),$$
accept a move from x to y. Set $X_{n+1} = y$. Otherwise, stay where you are an d set $X_{n+1} = x$.
3. Set $n = n + 1$ and return to Step 1.

In practice, we decide whether or not to accept the proposed move by drawing a $U_{n+1} \sim unif(0, 1)$ from a random number generator and then

- If $U_{n+1} \leq \alpha(x, y)$, set $X_{n+1} = y$.
- If $U_{n+1} > \alpha(x, y)$, set $X_{n+1} = x$.

After running this simulation forward "for a long time", output the result as an approximate draw from π. For truly iid draws, one would then start all over again. For approximate iid draws when computation is expensive, one would output values of a single chain "every so often" after an appropriate "burn-in time". Guidelines for all of these phrases in quotes are problem dependent.