36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 35: December 5

Lecturer: Siva Balakrishnan

35.1 Review and Outline

In the last class we were discussing sampling and integration:

1. Monte Carlo
2. Importance Weighting
3. Markov Chains

Today we will continue our discussion of Markov Chains and particularly try to understand the Metropolis Hastings algorithm.

35.2 Markov Chains

In the Bayesian problem we began with we cannot really use either Monte Carlo or importance weighting. We instead use MCMC.

First, the big picture: In Markov Chain Monte Carlo (MCMC) the goal is to design a Markov Chain, whose stationary distribution is the distribution P under which we want to compute an integral. We then sample from the Markov Chain, and then use the law of large numbers (adapted to Markov Chains) to argue that our estimate is good.

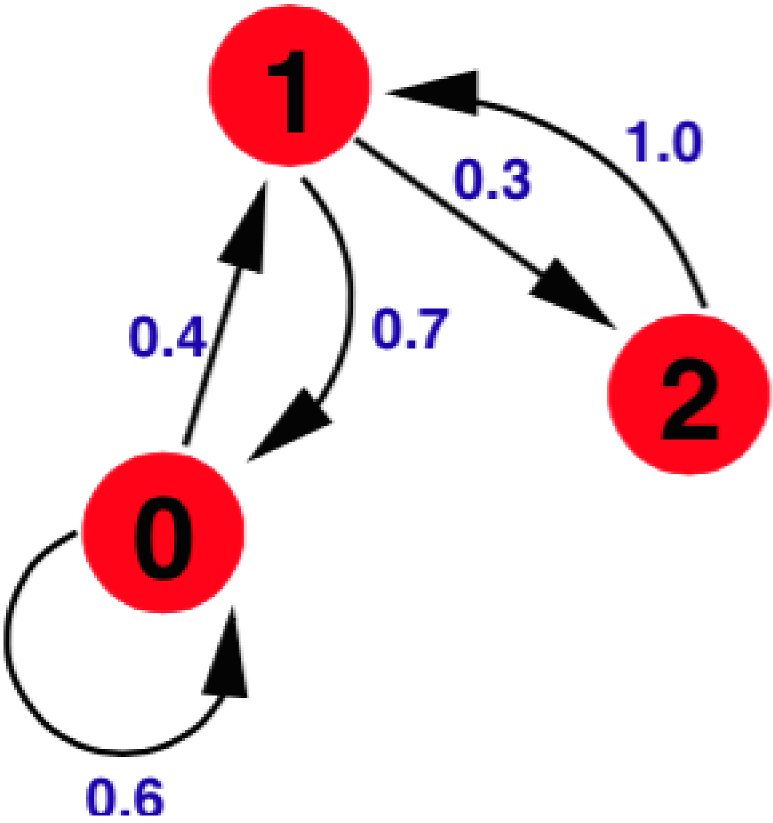
The details are a bit involved so we will get started today and finish up in the next lecture. We will do this at a very high-level.

First, what is a Markov Chain? A Markov Chain is a collection of random variables $\{X_1, \ldots, X_n\}$ that forms a graphical model: $X_1 \to X_2 \to \ldots \to X_n$.

In order to specify the joint distribution of a Markov Chain, we need to specify $p(X_1)$ and $p(X_{i+1}|X_i)$. To do this conveniently the focus is usually on what are called time-homogenous Markov Chains, i.e. $p(X_{i+1}|X_i) = T(X_i, X_{i+1})$, for a function T that does not depend on i. This function T is called the transition matrix or transition kernel of the Markov Chain.

Let us consider a simple example: suppose all the variables are discrete and take values $\{0, 1, 2\}$.

Consider a diagram of the form:



This diagram is specifying the transition matrix for us. Particularly, it says that the probability: $P(X_{i+1} = 2|X_i = 1) = 0.3$.

The next thing, we need to know is that Markov Chains almost always have a stationary distribution. The stationary distribution roughly, is the distribution of the random variable X_n for n large. We denote it by π, i.e.

$$\pi(x) = P(\lim_{n\to\infty} X_n = x).$$

An important aspect of Markov Chains is that they forget their initial state (i.e. they are ergodic/they mix), so π does not depend on $p(X_1)$.

Markov Chains also have a LLN associated with them: If $\{X_1, \ldots, X_n\}$ is a Markov Chain with stationary distribution π, then

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \int_x \pi(x) f(x) dx = \mathbb{E}_\pi[f(x)].$$

This is a pretty magical property. The Markov Chain samples are actually dependent, i.e. X_2 depends on X_1, and X_3 depends on X_2 and X_1, and so on. The LLN says that even though there are these dependencies, they are weak (and get much weaker as you get further away in the chain), so samples from a Markov Chain behave quite similarly to i.i.d. samples from the stationary distribution.

With all of these preliminaries in place only one thing remains: construct a Markov Chain with a particular stationary distribution (think, the posterior distribution).

35.3 Computing the stationary distribution

Suppose I specified a Markov chain and promised you that it had a well-defined (unique) stationary distribution. Then one basic question is how can I analytically compute what the stationary distribution is?

There are several ways to do this. One that we will use today is to check what are called *detailed balance* conditions, i.e. any distribution π (i.e. $\pi \geq 0$ and $\int \pi(x)dx = 1$), that satisfies:

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

for every x, y, where T is the transition kernel of the Markov chain, is the stationary distribution of the Markov chain. As a quick note: the reverse implication is not necessarily true, i.e. the stationary distribution of a Markov chain does not need to satisfy detailed balance (when it does, such Markov chains are called reversible Markov chains).

To summarize, if we construct a Markov chain, and there is some distribution π that satisfies the detailed balance conditions then that distribution is the stationary distribution of the Markov chain.

35.4 The Metropolis-Hastings algorithm

Recall, that our broad goal is to draw samples from a distribution f (say).

Choose X_0 arbitrarily. For each subsequent index i we follow the algorithm given below:

1. Sample a proposal $y \sim q(y|X_i = x)$ from a "proposal distribution" q.
2. Evaluate the ratio:

$$r = \min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\}.$$

3. Accept the new sample Y with probability r, and reject it otherwise. Alternatively, think of sampling $u \sim U[0,1]$ and accept if $u \leq r$ and reject otherwise.

Some basic intuition: We will understand formally why this works, but for now consider the case when the proposal is symmetric, i.e. $q(y|x) = q(x|y)$. In this case, we accept a new sample with probability:

$$r = \min\left\{\frac{f(y)}{f(x)}, 1\right\}.$$

Now, let us think about what it means to sample from f, roughly we want to draw more samples where f is high, and fewer samples where f is low. Our rule above basically says, always accept a sample if the density is higher at the proposed point (like hill-climbing) and if the density is lower at the proposed point you accept it with a smaller probability. This sampling rule is effectively biased to accept samples from regions where the density is high.

Three tasks remain: we need to decide how to choose a proposal distribution, we need to show that this algorithm does what we set out to, i.e. roughly generates samples from f, and we need to understand why this is useful in sampling from the posterior distribution for example.

35.4.1 Choosing a proposal distribution

This one is mostly an art, i.e. you try to pick a proposal distribution that somehow approximates the shape of the distribution you care about (f).

Often what we do is to choose:

$$q(Y|X = x) \sim N(x, \sigma^2),$$

so we sample a proposal around our current data point, and try to tune the tuning parameter σ (by trying to maintain a reasonable acceptance ratio while still enforcing that we explore most of the space).

35.4.2 Sampling posteriors

Our goal in the beginning of last lecture was to sample from the posterior distribution:

$$\pi(\theta|X_1, \ldots, X_n) = \frac{\pi(\theta)\mathcal{L}(\theta|X_1, \ldots, X_n)}{\int \pi(\theta)\mathcal{L}(\theta|X_1, \ldots, X_n)d\theta}.$$

More generally, suppose we have a distribution that we know up to the normalizing constant, i.e. we can compute $g(x)$ but we want to sample from f which is given by:

$$f(x) = \frac{g(x)}{\int g(x)dx},$$

and the denominator can be difficult to compute.

The key point: The Metropolis Hastings algorithm, only interacts with f through ratios of the form

$$\frac{f(x)}{f(y)} = \frac{g(x)}{g(y)},$$

which are easy to compute. When you take the ratio, the normalizing constant disappears.

35.4.3 Stationary distributions

Now, we go back to the Metropolis Hastings algorithm. We need to show that this algorithm is constructing a Markov chain and the stationary distribution of this Markov chain is f.

The first part is easy: each subsequent sample X_{i+1} only depends on X_i and q and does not depend on any of the prior $X_1, \ldots, X_{i-1}$ (conditional on X_i) so the samples $X_1, \ldots, X_n$ form a Markov chain.

Let us first understand the transition probabilities of our Markov chain. In order to transition from x to y we need to sample y from the proposal and then need to accept this proposal. This happens with probability:

$$T(x, y) = q(y|x) \min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\}.$$

Using this we can see that detailed balance is satisfied and f is the stationary distribution if:

$$f(x)q(y|x) \min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\} \overset{?}{=} f(y)q(x|y) \min\left\{\frac{f(x)q(y|x)}{f(y)q(x|y)}, 1\right\}.$$

This is easy to check by some case analysis. For instance, suppose $f(x)q(y|x) \geq f(y)q(x|y)$, then this reduces to:

$$f(x)q(y|x)\frac{f(y)q(x|y)}{f(x)q(y|x)} \overset{?}{=} f(y)q(x|y),$$

which is clearly true. We can similarly check this is true in the case when $f(x)q(y|x) < f(y)q(x|y)$. From this we can conclude that f is the stationary distribution of the Markov chain we have constructed.

35.4.4 Some caution

While MCMC is a really nice trick in order to generate samples from something close to the posterior, there is an important caveat that I have ignored. For a Markov chain, the stationary distribution is its "asymptotic distribution", i.e. it is the distribution you are getting samples from asymptotically (as $n \to \infty$).

The hope is usually that for small (finite) values of n the distribution is close to the stationary distribution. This is called mixing or rapid mixing. Unfortunately, however, in many cases we do not know if the Markov chain mixes rapidly (this depends in a complicated fashion on the proposal and the unknown density f).

To a large extent, MCMC is a sensible heuristic, and some caution/care is required in applying it to difficult problems.