MSA101/MVE187 2017 Lecture 5

Petter Mostad

Chalmers University

September 12, 2017

Importance sampling

- MC integration computes

$$\int h(x)f(x)\,dx$$

where $f(x)$ is a probability density function, by simulating $x_1, \ldots, x_m$ according to f and taking the averages of $h(x_1), \ldots, h(x_m)$. The result has accuracy $\sqrt{Var(h(X))/m}$.

- Instead, we may re-write the integral as

$$\int \left[\frac{h(x)f(x)}{g(x)}\right] g(x)\,dx$$

and simulate x_i according to g and taking the averages of $h(x_1)f(x_1)/g(x_1), \ldots, h(x_m)f(x_m)/g(x_m)$.

- A good idea if $Var(h(X)f(X)/g(X))$ is much smaller than $Var(h(X))$.

Sampling importance resampling

- The similar idea to importance sampling, but now used to obtain an approximate sample from the target distribution.
- The algorithm is: Sample $X_1, \ldots, X_n$ from the g density, then resample from these (with replacement) using weights

$$w_i = \frac{f(X_i)/g(X_i)}{\sum_{j=1}^{n} f(X_j)/g(X_j)}$$

- The normalization of the weights produces a (usually small) bias.

MCMC simulation

General idea of Markov chain Monte Carlo:

- Construct a Markov chain which has as its *stationary distribution* the target distribution (the posterior) and simulate from this chain.
- From the simulations, extract something that is *approximately* a sample from the posterior.
- Do Monte Carlo integration with this sample.

Review of Markov chains

- Definition: A (discrete time, time-homogeneous) Markov chain with kernel K is a sequence of random variables $X^{(0)}, X^{(1)}, X^{(2)}, \dots$ satisfying, for all t,

$$\pi(X^{(t)} \mid X^{(0)}, X^{(1)}, \dots, X^{(t-1)}) = \pi(X^{(t)} \mid X^{(t-1)}) = K(X^{(t-1)}, X^{(t)})$$

- A stationary distribution f is one satisfying

$$f(y) = \int K(x, y) f(x)\, dx$$

- Example: In the case of a state space with n possible values, a distribution is represented by a vector of length n summing to 1, and K is represented by an $(n \times n)$ matrix with rows summing to 1. A stationary distribution is a (left) eigenvector for K.

Conditions for existence of a *unique* stationary distribution

- Rerducibility / irreducible
- Periodicity / aperiodic
- Transience / recurrent
- Ergodic / ergodicity
- In an irreducible, aperiodic, recurrent chain, $X^{(n)}$ converges to a unique stationary distribution when $n \to \infty$.

The detailed balance condition

- A Markov chain satisfies the *detailed balance condition* relative to a density f if, for all x, y,

  $$f(x)K(x,y) = f(y)K(y,x)$$

  where $K(x,y)$ is the kernel of the Markov chain. Called a *reversible* Markov chain.
- If a chain satisfies detailed balance relative to f, then f must be a stationary distribution.
- Prove by integrating over x!

The Metropolis-Hastings algorithm

Given a probability density f that we want to simulate from. Construct a *proposal function* $q(y \mid x)$ which for every x gives a probability density for a proposed new value y. The algorithm starts with a choice of an initial value $x^{(0)}$ for x, and then simulates $x^{(t)}$ given $x^{(t-1)}$. Specifically, given $x^{(t)}$,

- Simulate a new value y according to $q(y \mid x^{(t)})$.
- Compute the acceptance probability

$$\rho(x^{(t)}, y) = \min\left(\frac{f(y)q(x^{(t)} \mid y)}{f(x^{(t)})q(y \mid x^{(t)})}, 1\right).$$

- Set

$$x^{(t+1)} = \begin{cases} y & \text{with probability } \rho(x^{(t)}, y) \\ x^{(t)} & \text{with probability } 1 - \rho(x^{(t)}, y) \end{cases}$$

The chain defined by Metropolis-Hastings satisfies the detailed balance condition

- Assume first that $\rho(x,y) < 1$ (with $x \neq y$). Then

$$\begin{aligned} f(x)K(x,y) &= f(x)q(y \mid x)\rho(x,y) = f(x)q(y \mid x)\frac{f(y)q(x \mid y)}{f(x)q(y \mid x)} \\ &= f(y)q(x \mid y) = f(y)q(x \mid y)\rho(y,x) = f(y)K(y,x) \end{aligned}$$

  The next to last step is because $\rho(y,x) = 1$ when $\rho(x,y) < 1$.
- If we start with $\rho(x,y) = 1$ the situation is clearly symmetrical, and we get the same result.

The Ergodic theorem

- This theorem says that, when $X^{(0)}, \ldots, X^{(t)}, \ldots,$ is sampled from an ergodic Markov chain with stationary distribution f, we have that

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} h(X^{(t)}) = E_f[h(X)]$$

- When the sample is instead a random sample from f, this is the law of large numbers; we then also have the extension to the Central Limit Theorem, telling us how fast the convergence is.
- In the ergodic case, we still have convergence, but we don't know as easily how fast it is.

Note that...

- ...the Metropolis-Hastings algorithm *only* requires knowledge of the target density $f(x)$ up to a constant not involving x, as the density only appears in the quotient $f(y)/f(x)$ in the algoritm.
- ...the Metropolis-Hastings algorith *only* requires knowledge of the proposal density up to a constant, for the same reason.
- ...similarly, smart versions of the Metropolis-Hastings algorithm uses proposal flunctions so that many factors in the acceptance probability
  $$\frac{f(y)q(x \mid y)}{f(x)q(y \mid x)}$$
  cancel each other.

Example: Symmetric proposal functions

Random walk Metropolis-Hastings

- We use

$$q(y \mid x) = g(y - x), \text{where } g(-x) = g(x) \text{ for all } x.$$

  for some density function g: The proposal becomes symmetric around x

- This means that $q(y \mid x) = q(x \mid y)$ and the acceptance probability becomes

$$\min(\frac{f(y)}{f(x)}, 1)$$

  where f is the target density.

- Example: $y = x + \epsilon$, where $\epsilon \sim \text{Normal}(0, \Sigma)$ for some covariance matrix Σ.
- The scaling of the size of the jumps can be very trickiy to get right, to produce good convergence of the Markov chain.

Example: Independent proposal functions

- A simple special case is when $q(y \mid x)$ does not depend on x; i.e. proposals are independently generated from $q(y)$.
- The generated values are however *not* independent: When the proposed value is not accepted, the new value in the chan is equal to the old.
- Note that, if the ratio $f(x)/q(x)$ is unbounded, the chain can become stuck in such point where this ratio is too high. Then the convergence can be very bad.

Gibbs sampling

- The idea: Sampling from conditional distributions $\pi(X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$ for the target density. These are in many cases easy to derive.
- Two stage and multistage Gibbs sampling.
- Why does it work? Easy to show that the Markov chain satisfies the detailed balance condition.
- Examples RC 7.1, 7.2
- Example RC 7.3: Simulating from a posterior that does not have an analytic form, but where each of the conditional distributions has an analytic form.

Example

- An example: A bivariate Normal distribution multiplied with the indicator function for some convex set.
- Can you think of alternative methods of simulating from the distribution?
- Consider a bivariate Normal distribution multiplied with the indicator function for the set $[0, 1] \times [0, 1] \cup [2, 3] \times [2, 3]$. How does Gibbs sampling work now?
- Can you think of various ways to simulate from the distribution above?