MVE550 2019 Lecture 8

Petter Mostad

Chalmers University

December 3, 2019

Overview

- Overview of MCMC.
- Strong law of large numbers for ergodic Markov chains.
- The Metropolis Hastings algorithm. Example.
- Gibbs sampling. Example.

Markov chain Monte Carlo (MCMC)

- A hugely important and useful set of algorithms.
- In particular useful in Bayesian statistics.
- NOTE: So far, we have used Bayesian inference to learn about, for example, parameters in Markov chains. Now, we use Markov chains as a tool to do Bayesian statistics.

Laws of large numbers

- Strong law of large numbers for samples: If $Y_1, Y_2, \ldots, Y_m$ and Y are independent random variables from a distribution with finite mean, and if r is a bounded function, then, with probability 1,

$$\lim_{m\to\infty} \frac{r(Y_1) + r(Y_2) + \cdots + r(Y_m)}{m} = \mathrm{E}[r(Y)]$$

- Strong law of large numbers for Markov chains: If $X_0, X_1, \ldots,$ is an ergodic Markov chain with stationary distribution π, and if r is a bounded function, then, with probability 1,

$$\lim_{m\to\infty} \frac{r(X_1) + r(X_2) + \cdots + r(X_m)}{m} = \mathrm{E}[r(X)]$$

  where X has the stationary distribution π.
- Note that this holds not only for Markov chains with discrete state spaces, but also for Markov chains with continuous distributions (which we will look at later).
- NOTE: When using this theorem in practice, one might improve accuracy by throwing away the first sequence $X_1, \ldots, X_s$ for $s < m$ before computing the average. This first sequence is called the *burn-in*.

Overview: MCMC for Bayesian inference

- We assume we know the densities for the likelihood $\pi(y \mid \theta)$ and the prior $\pi(\theta)$. Then $\pi(\theta \mid y) \propto_\theta \pi(y \mid \theta)\pi(\theta)$ gives us the posterior density *up to a constant*.
- If we have a sample $\theta_1, \ldots, \theta_m$ from the posterior $\theta \mid y$, we can approximate predictions by the strong law of large numbers as follows:

$$\pi(y_{new} \mid y) = \int \pi(y_{new} \mid \theta, y)\pi(\theta \mid y)\, d\theta \approx \frac{1}{m}\sum_{i=1}^{k} \pi(y_{new} \mid \theta_i)$$

- MCMC does not provide a sample, but instead a sequence $\theta_1, \ldots, \theta_m$ so that the above holds when $m \to \infty$ by the strong law of large numbers for ergodic Markov chains.
- NOTE: More generally, to find the expected value of a variable $x = f(\theta)$ under the posterior distribution, we can use

$$\mathrm{E}[x] = \int f(\theta)\pi(\theta \mid y)\, d\theta \approx \frac{1}{m}\sum_{i=1}^{m} f(\theta_i).$$

The Metropolis Hastings algorithm

- Assume a density (or probability mass function) $\pi(\theta)$ is provided.
- We also assume given a *proposal function* $q(\theta_{new} \mid \theta)$, which, for every given θ, provides a probability distribution (or probability mass function) for a new θ_{new}.
- Finally, define, for θ and θ_{new}, the acceptance probability

$$a = \min\left(1, \frac{\pi(\theta_{new})q(\theta \mid \theta_{new})}{\pi(\theta)q(\theta_{new} \mid \theta)}\right)$$

- The Metropolis Hastings algorithm is: Starting with some initial value θ_0, generate $\theta_1, \theta_2, \ldots$ by, at each step, proposing a new θ based on the old using the proposal function and accepting it with probability a. If it is not accepted, the old value is used again.
- If this defines an ergodic Markov chain, its unique stationary distribution is $\pi(\theta)$ (Proof below).

The Metropolis Hastings algorithm, continued

NOTE:

- The computations for good binary sequences is an example of this, with $\pi(\theta)$ uniform and q the random walk.
- The density $\pi(\theta)$ only needs to be known up to a constant.
- If the proposal function is symmetric, i.e., $q(\theta \mid \theta_{new}) = q(\theta_{new} \mid \theta)$ for all θ and θ_{new}, then q disappears in the formula for the acceptance probaility a.
- Unless the distribution $\pi(\theta)$ is *positive*, remark 4 in Dobrow page 188 does NOT hold. If $\pi(\theta)$ is not positive, ergodicity of the Metropolis Hastings Markov chain needs to be checked separately, even if the proposal Markov chain is ergodic.

Proof that MH algorithm works

- In fact, we will show that the Metropolis Hastings chain fulfills the detailed balance condition relative to $\pi(\theta)$. Thus it is time reversible and if it is ergodic it will have $\pi(\theta)$ as its limiting distribution.
- Let $T(\theta_{i+1} \mid \theta_i)$ be the transition function for the MH Markov chain. Assume $\theta_{i+1} \neq \theta_i$, and

  $$\frac{\pi(\theta_{i+1})q(\theta_i \mid \theta_{i+1})}{\pi(\theta_i)q(\theta_{i+1} \mid \theta_i)} \leq 1$$

  Then

  $$\begin{aligned} \pi(\theta_i)T(\theta_{i+1} \mid \theta_i) &= \pi(\theta_i)q(\theta_{i+1} \mid \theta_i)\frac{\pi(\theta_{i+1})q(\theta_i \mid \theta_{i+1})}{\pi(\theta_i)q(\theta_{i+1} \mid \theta_i)} \\ &= \pi(\theta_{i+1})q(\theta_i \mid \theta_{i+1}) = \pi(\theta_{i+1})T(\theta_i \mid \theta_{i+1}) \end{aligned}$$

- We get a similar computation when the opposite inequality holds.

More notes on the MH algorithm

- We have so far worked with Markov chains where the state space is discrete. However, the theory we need for the Metropolis Hastings method to work is unchanged also if the state space is continuous, or even multivariate with a mix of continuous and discrete variables.
- Note that the proposal distribution can be chosen with almost total freedom (as long as one can prove that the resulting MH Markov chain becomes ergodic). The choice of proposal function generally has a large influence on the rate of convergence of the MH chain, and thus on the accuracy of results!
- If the target density is positive on the same set as the the one where the proposal function generates proposals, and if the proposal function is ergodic, then the MH chain is ergodic.

Example

- Assume that a model has the real parameter θ, and that the posterior for θ has been found to be

  $$\pi(\theta \mid \text{data}) = 0.3\,\text{Normal}(\theta; 2, 0.5^2) + 0.7\,\text{Normal}(\theta; 6, 1^2).$$

  As a test example, compare a sample simulated directly from this distribution to one simulated using Metropolis Hastings. Use as starting value 1 and proposal function $\pi(\theta' \mid \theta) = \text{Uniform}(\theta'; \theta - 0.5, \theta + 0.5)$.
- Assume we would like find the predictive distribution for y when $y \mid \theta \sim \text{Normal}(\theta, 0.3^2)$ and θ has the distribution above.
  - Do this first by using a sample from generated by Metropolis Hastings.
  - Then, compute and compare to the theoretical distribution.