MVE550 2023 Lecture 10
Perfect sampling
More on MCMC (review)

Petter Mostad

Chalmers University

November 24, 2023

Gibbs sampling

- For any probability model over a vector $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, consider a MH proposal function changing only one coordinate, with the value of this coordinate simulated from the conditional distribution given the remaining coordinates.
- Prove that the acceptance probability is 1.
- Putting together an algorithm updating different coordinates in different steps may create an ergodic Markov chain.
- This is then called *Gibbs sampling*.
- Sometimes the conditional distributions are easy to derive. Then this is an easy-to-use version of Metropolis Hastings.

The Ising model

- Uses a grid of vertices; we will assume an $n \times n$ grid. Two vertices v and w are *neighbours*, denoted $v \sim w$, if they are next to each other in the grid.
- Each vertex v can have value $+1$ or -1 (called its "spin"); we denote this by $\sigma_v = 1$ or $\sigma_v = -1$.
- A *configuration* σ consists of a choice of $+1$ or -1 for each vertex: Thus the set Ω of possible configurations has $2^{(n^2)}$ elements.
- We define the *energy* of a configuration as $E(\sigma) = -\sum_{v \sim w} \sigma_v \sigma_w$.
- The Gibbs distribution is the probability mass function on Ω defined by

$$\pi(\sigma) \propto_\sigma \exp\left(-\beta E(\sigma)\right)$$

  where β is a parameter of the model; $1/\beta$ is called the *temperature*.
- It turns out that when the temperature is high, samples from the model will show a chaotic pattern of spins, but when the temperature sinks below the *phase transition* value, in our case $1/\beta = 2/\log(1+\sqrt{2})$, samples will show chunks of neighbouring vertices with the same spin; the system will be "magnetized".

Simulating from the Ising model using Gibbs sampling

- For a vertex configuration σ and a vertex v let σ_{-v} denote the part of σ that does not involve v.
- Propose a new configuration σ^* given an old configuration σ by first choosing a vertex v, then, let σ^* be identical to σ except possibly at v: Decide the spin at v using the conditional distribution given σ_{-v}:

$$\pi(\sigma_v = 1 \mid \sigma_{-v}) = \frac{\pi(\sigma_v = 1, \sigma_{-v})}{\pi(\sigma_{-v})} = \frac{\pi(\sigma_v = 1, \sigma_{-v})}{\pi(\sigma_v = 1, \sigma_{-v}) + \pi(\sigma_v = -1, \sigma_{-v})}$$

$$\begin{aligned}
&= \frac{1}{1 + \frac{\pi(\sigma_v = -1, \sigma_{-v})}{\pi(\sigma_v = 1, \sigma_{-v})}} = \frac{1}{1 + \exp\left(-\beta E(\sigma_v = -1, \sigma_{-v}) + \beta E(\sigma_v = 1, \sigma_{-v})\right)} \\
&= \frac{1}{1 + \exp\left(\beta \sum_{v \sim w} \sigma_v \sigma_w \mid_{\sigma_v = -1} - \beta \sum_{v \sim w} \sigma_v \sigma_w \mid_{\sigma_v = 1}\right)} \\
&= \frac{1}{1 + \exp\left(-2\beta \sum_{v \sim w} \sigma_w\right)}.
\end{aligned}$$

- This works. However, we will see below an even better approach, "perfect sampling", to the Ising model simulation problem.

Reminder: The Metropolis Hastings algorithm

- Goal: Given $f(\theta)$ proportional to some probability (density) function $\pi(\theta)$, simulate from a Markov chain whose limiting distribution is $\pi(\theta)$, apply a function to the simulated values and average, to make approximate inference.
- To simulate, we need a *proposal distribution* $q(\theta_{new} \mid \theta)$, which, for every given θ, provides a probability (density) function for a θ_{new}.
- At each Markov step, simulate a proposal, and accept it with probability

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

  or else repeat the old value.
- *The main problem with MCMC*: Difficult to know the connection between the length of the sample and the accuracy of inference results.

Knowing convergence has been reached: Perfect sampling

Given ergodic Markov chain with finite sample space of size k and limiting distribution π.

- Idea: Given n, prove that X_n actually has reached the limit distribution.
- Method: Prove that the distribution at X_n is independent of the starting value at X_0.
- Try: Construct k Markov chains that are dependent ("coupled") but which are marginally Markov chains as above. If they start at the k possible values at X_0 but have identical values at X_n, we are done.
- Note: *n cannot* be determined as the first value where the k chains meet; it must be determined independently of such information!
- Thus usually one wants to generate chains $X_{-n}, X_{-n+1}, \ldots, X_0$ where X_0 has the limiting distribution, and we stepwise increase n to make all chains *coalesce* to one chain.

Using same source of randomness for all k chains

Consider the chains $X_{-n}^{(j)}, \ldots, X_0^{(j)}$ for $j = 1, \ldots, k$.

- Instead of simulating $X_{i+1}^{(j)}$ based on $X_i^{(j)}$ independently for each j, we define a function g so that $X_{i+1}^{(j)} = g(X_i^{(j)}, U_i)$ for all j, where $U_i \sim \text{Uniform}(0, 1)$.
- Thus if two chains have identical values in X_i, they will also be identical at X_{i+1}.
- See Figure 5.10 in Dobrow.
- Thus, for a particular n, if all chains have not converged at X_0, we simulate k chains from X_{-2n} to X_{-n}: They might only hit a subset of the k states at X_{-n} and thus might coalesce to one state at X_0, using the old simulations. If not, double n again.

Monotonicity

- Do we need to keep track of *all* k chains?
- We define a *partial ordering* on a set as a relation $x \leq y$ between *some* pairs x and y in the set, such that:
  - If $x \leq y$ and $y \leq x$ then $x = y$.
  - If $x \leq y$ and $y \leq z$ then $x \leq z$ (in fact we don't need this).
- We will need that our partial ordering has a minimal element (an m such that $m \leq x$ for all x) and a maximal element (an M such that $x \leq M$ for all x).
- If we have a partial ordering on the state space of the Markov chain, and if $x \leq y$ implies $g(x, U) \leq g(y, U)$, then g is *monotone*.
- We can then prove that we only need to keep track of the chain starting at m and the chain starting at M!

Example: Perfect simulation from the Ising model

- Given an Ising model with $\beta > 0$.
- Define partial ordering on Ω (the set of all configurations) as follows

$$\sigma \leq \tau \text{ if } \sigma_v \leq \tau_v \text{ for all vertices } v$$

- We have a minimal and a maximal configuration (all -1's and +1's, respectively).
- We can arrange for g, the updating of chains, to be monotone: Assuming $\sigma \leq \tau$,

$$\Pr(\sigma_v = 1 \mid \sigma_{-v}) = \frac{1}{1 + \exp(-2\beta \sum_{v \sim w} \sigma_w)} \leq \frac{1}{1 + \exp(-2\beta \sum_{v \sim w} \tau_w)} = \Pr(\tau_v = 1 \mid \tau_{-v}).$$

- So perfect simulation from the Ising model proceeds as follows: Start one chain m at all -1's and one chain M at all +1's. Cycle through the vertices and compute the conditional probabilities p_m and p_M of +1 at that vertex. We know that $p_m \leq p_M$. Simulate $U \sim \text{Uniform}(0, 1)$. If $U < p_m$ set $\sigma_v = -1$ for both chains, and if $U > p_M$ set $\sigma_v = +1$ for both chains. Otherwise set $\sigma_v = +1$ for the M chain and $\sigma_v = -1$ for the m chain. Determine coalescence as above.

From lecture 8: Second example

- We have observed the data (x_i, y_i):

$$(2, 0.32), (3, 0.57), (4, 0.61), (6, 0.83), (9, 0.91)$$

- The context gives us the following model
  - We expect the data to follow $y = f(x, \theta_1) = \frac{\exp(\theta_1 x) - 1}{\exp(\theta_1 x) + 1}$ where θ_1 is an unknown positive parameter.
  - We have observed the data with added noise $\text{Normal}(0, \theta_2^2)$ where θ_2 is an unknown positive parameter.
  - We assume a flat prior on $\theta_1 > 0$ and $\theta_2 > 0$.
- We get the posterior

$$\pi(\theta \mid \text{data}) \propto_\theta \prod_{i=1}^{5} \text{Normal}(y_i; f(x_i, \theta_1), \theta_2^2).$$

- Use MCMC to simulate from the value of y when $x = 10$ (see R).