MVE550 2020 Lecture 7.1
Dobrow Sections 5.1, 5.2
The Metropolis Hastings algorithm

Petter Mostad

Chalmers University

November 21, 2020

The limiting distribution as target distribution

- So far: Start with a Markov chain, learn what happens when the number of steps approaches ∞.
- We now turn this on its head: Start with defining a limiting distribution, call it the "target distribution", then derive a Markov chain with this limiting distribution.
- Purpose: If we sample the Markov chain for sufficiently many steps, we know that we have an approximate sample from our target distribution.
- This is useful in situations where we need a sample, but sampling directly is difficult.

Is an approximate sample good enough?

- Strong law of large numbers for samples: If $Y_1, Y_2, \ldots, Y_m$ and Y are i.i.d. 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 of continuous random variables (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*.

Toy example

- Consider the Markov chain $X_0, X_1, \ldots$ with states $\{0, 1, 2\}$ and with

$$P = \begin{bmatrix} 0.99 & 0.01 & 0 \\ 0 & 0.9 & 0.1 \\ 0.2 & 0 & 0.8 \end{bmatrix}.$$

  Using theory from Chapter 3 we get that the limiting distribution is $v = (20/23, 2/23, 1/23)$.
- Consider the function $r(x) = x^5$. If X is a random variable with the limiting distribution,

$$\mathrm{E}\left(r(X)\right) = 0^5 \cdot \frac{20}{23} + 1^5 \cdot \frac{2}{23} + 2^5 \cdot \frac{1}{23} = \frac{33}{23} = 1.4348$$

- If $Y_1, \ldots, Y_n$ are all i.i.d. variables with the limiting distribution, we can check numerically (see R code) that

$$\lim_{n\to\infty} \frac{r(Y_1) + \cdots + r(Y_n)}{n} = 1.4348$$

- We also get (see R code), for $X_0, X_1, \ldots,$ that

$$\lim_{n\to\infty} \frac{r(X_1) + \cdots + r(X_n)}{n} = 1.4348$$

  but in this case the limit is approached more slowly.

Less toy-ish example: "Good" sequences

Consider sequences of length m consisting of 0's and 1's.

- A sequence is called "good" if if contains no consecutive 1's.
- What is the average number of 1's in good sequences of length m?
- Direct computation is possible, but not obvious how to do.
- *Efficient* direct simulation of a sample of good seqences is not obvious how to do, when m is, say, above 100.
- We construct a random walk on a weighted graph with nodes consisting of all good sequences (fixed m).
  - The limiting distribution is the uniform distribution.
  - Thus we can estimate the solution by counting 1's in sequences generated by the Markov chain, and then take the average.
  - This is both easy to program and gives efficient and accurate results.
- Construct the graph as follows:
  - Two good sequences are neighbours when the differ at exactly one position. The weight of edge connecting them is 1.
  - Each good sequence has an edge connecting it to itself, with weight so that the total weights of edges going out from the sequence is m.

The Metropolis Hastings algorithm

If we start with a particular distribution, can we construct a Markov chain with that as the limiting distribution?

- Let θ be a discrete random variable with probability mass function $\pi(\theta)$.
- We also assume given a *proposal distribution* $q(\theta_{new} \mid \theta)$, which, for every given θ, provides a 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

NOTES:

- The computations for good 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

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

  the last step because, with assumption above, $\frac{\pi(\theta_i)q(\theta_{i+1}|\theta_i)}{\pi(\theta_{i+1})q(\theta_i|\theta_{i+1})} \geq 1$
- We get a similar computation when the opposite inequality holds.

Example 1: Cryptography (from Dobrow)

- A simple way to encrypt a text is to apply to each character a fixed permutation f of the set of the 26 English characters plus space. The text can be decrypted by applying the reverse permutation, if it is known. If T is a text we write $f(T)$ for T encrypted with f.
- Given a short encrypted text $f(T)$, can we find the permutation f?
- Using a text database we first fit a Markov model for text by counting transitions between consecutive characters.
- For any text T', we can then compute the probability $S(T')$ for T' being observed as a sequence in this Markov model.
- We get a probability distribution on the set of all the permutations above by defining, for any f,

$$\pi(f) \propto_f S(f(T))$$

- We use Metropolis Hastings with a proposal function that picks two characters at random and adds to f a switch of these.
- The density $\pi(f)$ is on a very large set, with very few of the f having significant probability. Yet M.H. manages to find these (or this) f.

Example 2: Darwin's finches (from Dobrow)

- A *co-occurrence matrix* M has different species as rows and different locations as columns. If a species occurs at a location, the matrix contains 1, otherwise 0.
- A *checkerboard* is a submatrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ or $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Let $C(M)$ count the number of checkerboards in M .
- Darwin made a co-occurrence matrix for finches on the Galapagos islands. Compared to the set Ω of possible co-occurrence matrices with the same marginal sums, did it contain an unusually large number of checkerboards?
- Use Metropolis Hastings to simulate from the uniform distribution on Ω. Use a proposal function that uniformly randomly locates one of the checkerboards and switches it to the opposite form.
- The acceptance probability becomes $\min(1, C(M)/C(M^*))$ where M^* is proposed from M (error in Dobrow!)
- Simulation results show that the number of checkerboards observed by Darwin (333) is indeed unexpectedly large, proving competition between the finches.

MVE550 2020 Lecture 7.2
Dobrow Sections 5.3, 5.4
Ising model. Gibbs sampler. Perfect sampling

Petter Mostad

Chalmers University

November 21, 2020

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 density 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 Metropolis Hastings

- 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}:

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

$$= \frac{1}{1 + \frac{\pi(\sigma_v = -1, \sigma_{-v})}{\pi(\sigma_v = 1, \sigma_{-v})}} = \frac{1}{1 + \exp\left(-\beta E(\sigma_v = -1, \sigma_{-v}) + \beta E(\sigma_v = 1, \sigma_{-v})\right)}$$

$$= \frac{1}{1 + \exp\left(\beta \sum_{v \sim w} \sigma_v \sigma_w \mid_{\sigma_v = -1} - \beta \sum_{v \sim w} \sigma_v \sigma_w \mid_{\sigma_v = 1}\right)}$$

$$= \frac{1}{1 + \exp\left(-2\beta \sum_{v \sim w} \sigma_w\right)}.$$

- As $\sigma_{-v} = \sigma^*_{-v}$ we get $\frac{\pi(\sigma^*) q(\sigma | \sigma^*)}{\pi(\sigma) q(\sigma^* | \sigma)} = \frac{\pi(\sigma^*_v | \sigma^*_{-v}) \pi(\sigma^*_{-v}) \pi(\sigma_v | \sigma^*_{-v})}{\pi(\sigma_v | \sigma_{-v}) \pi(\sigma_{-v}) \pi(\sigma^*_v | \sigma_{-v})} = 1$ so the acceptance probability is always 1!

Gibbs sampling

- In the Ising model, the states can be written as a vector $\sigma = (\sigma_1, \ldots, \sigma_{n^2})$ of components or coordinates. We used a proposal function which changed only one coordinate and simulated its new value using the conditional distribution given the remaining coordinates.
- For any probability model over a vector $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ we can do the same: The proposal function changes only one coordinate, and the distribution of this coordinate is given as the conditional distribution given the remaining coordinates. The proof that the acceptance probability is 1 is unchanged!
- This is called Gibbs sampling.
- Note that we may choose the coordinate to change in various ways, as long as the resulting Markov chain becomes ergodic.
- In the Ising model, the conditional distributions $\pi(\theta_k \mid \theta_{-k})$ are easy to derive and simulate from, and this may often be the case. In such cases, Gibbs sampling is an easy-to-use version of Metropolis Hastings.

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.
- How: 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 beforehand!
- 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.
- If, for a particular n, 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.