Chapter 8

Markov chain Monte Carlo

8.1 The key idea of MCMC

We start with a state space S and a probability density $\pi(x)$ on it. Our goal is to come up with a Markov chain on this state space that has $\pi(x)$ as its invariant distribution. If the chain is recurrent, then the ergodic theorem says that we can compute (approximately) the expected value of a function $F(x)$ on the state space by running the chain for along time and taking the long time average of $F(x)$ along the sequence of states that we generate. We start with two very simple examples to illustrate the idea of MCMC. One is discrete, one continuous.

Example: Fix an integer k and let S be the set of permutations with on $\{1, 2, \cdots, k\}$. Let π be the uniform measure on S. We want to construct a Markov chain on S with π as the stationary measure. (There are many ways to do this.) Our algorithm is as follows. We pick two integers $i, j \in \{1, 2, \cdots, k\}$. The choice is random with the uniform distribution on the set of k^2 possibilities. Let σ_{ij} be the permutation that interchanges i and j and leaves the other elements fixed. Then if σ is the state at time n, the state at time $n + 1$ is $\sigma_{ij} \circ \sigma$.

Show that it satisfies detailed balance.

Show it is irreducible.

Remark: This example illustrates the following observation. If $p(x, y)$ is symmetric, i.e., $p(x, y) = p(y, x)$, then the stationary distribution is the uniform distribution.

Example: The state space is the real line. Let $\pi(x)$ be the density of the standard normal.

We want to cook up a Markov chain with this as the stationary distribution. We take

$$k(x,y) = c(\sigma)\exp(-\frac{1}{2}(y-\frac{1}{2}x)^2/\sigma^2) \tag{8.1}$$

Show that $\pi(x)$ is the stationary distribution if $\sigma^2 = 3/4$.

8.2 The Metropolis-Hasting algorithm

We want to generate samples from a distribution

$$\pi(x) = \frac{1}{Z}p(x) \tag{8.2}$$

where $x \in X$. The set X could be a subset of R^d in which case $\pi(x)$ is a density and the measure we want to sample from is $\pi(x)$ times Lebesgue measure on R^d. Or X could be finite or countable in which case the distribution is discrete and the measure we want to sample from assigns probability $\pi(x)$ to x. The function $p(x)$ is known and Z is a constant which normalizes it to make it a probability distribution. Z may be unknown.

Let $q(x,y)$ be some transition function for a Markov chain with state space S. If S is discrete then $q(x,y)$ is a transition probability, while if S is continuous it is a transition probability density. We will refer to q as the proposal density or distribution. $q(x,y)$ is often written as $q(y|x)$. We assume that $q(x,y) = 0$ if and only if $q(y,x) = 0$ We define a function called the acceptance probability by

$$\alpha(x,y) = \min\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\} \tag{8.3}$$

Since $\pi(y)/\pi(x) = p(y)/p(x)$, the possibly unknown constant Z is not needed to compute the acceptance probability. Note that $\alpha(x,y)$ is always in $[0,1]$. If one of the terms in the denominator above is zero, we define $\alpha(x,y)$ to be zero. It really doesn't matter how we define $\alpha(x,y)$ in this case. Explain.

Then we define a Monte Carlo chain as follows.

Metropolis-Hasting algorithm *Suppose the chain is in state X_n at time n. We generate Y from the distribution $q(y|X_n)$. Next generate U from the uniform distribution on $[0,1]$. If $U \leq \alpha(X_n, Y)$ then we set $X_{n+1} = Y$. Otherwise we set $X_{n+1} = X_n$.*

If the proposal distribution is symmetric, meaning that $q(y,x) = q(x,y)$, then the acceptance probability function simplifies to

$$\alpha(x,y) = \min\{\frac{\pi(y)}{\pi(x)}, 1\} \tag{8.4}$$

This is sometimes called just the Metropolis algorithm. This case was studied by Metropolis. Hasting generalized to non-symmetric q.

The above description of the algorithm implicitly defines the transition kernel for the Markov chain. We make it more explicit. In the discrete case for $x \neq y$ the transition probability is

$$p(x, y) = q(x, y)\alpha(x, y) \tag{8.5}$$

and

$$p(x, x) = q(x, x)\alpha(x, x) + \sum_y [1 - \alpha(x, y)]q(x, y) \tag{8.6}$$

The first term comes from accepting the proposed state x and the second term (with the sum on y) comes from proposing y and rejecting it.

In the continuous case the transition kernel $K(x, \cdot)$ is a mixture of a continuous measure and a point mass.

$$K(x, A) = \int_A \alpha(x, y)q(x, y)\, dy + 1_{x\in A} \int_S [1 - \alpha(x, y)]q(x, y)\, dy \tag{8.7}$$

Remark: This sort of looks like acceptance-rejection. We generate a proposed state Y and accept it with probability $\alpha(X_n, Y)$, reject it otherwise. But it is not the same as the acceptance-rejection algorithm. One crucial difference is that in the acceptance-rejection algorithm when we reject a proposed value the number of samples does not increase. In Metropolis-Hasting when we reject Y the chain still takes a time step. When this happens there are two consecutive states in $X_0, X_1, X_2, \cdots$ that are the same. So in the time average

$$\frac{1}{n} \sum_{k=1}^{n} f(X_k) \tag{8.8}$$

there can be terms that are the same.

Theorem 1 *$\pi(x)$ is the stationary distribution of the Markov chain of the Metropolis-Hasting algorithm.*

Proof: We will eventually consider the discrete and continuous cases separately, but first we prove the following crucial identity which holds in both cases.

$$\pi(x)\alpha(x, y)q(x, y) = \pi(y)\alpha(y, x)q(y, x), \quad \forall x, y \in S \tag{8.9}$$

To prove it we consider two cases: $\alpha(x,y) = 1$ and $\alpha(y,x) = 1$. (It is possible these cases overlap, but that does not affect the proof.) The cases are identical. So we assume $\alpha(x,y) = 1$. Then

$$\alpha(y,x) = \frac{\pi(x)q(x,y)}{\pi(y)q(y,x)} \tag{8.10}$$

The above identity follows.

Now consider the discrete case, i.e., the state space is finite or countable. In this case $\pi(x)$ is a probability mass function. The transition function is

$$p(x,y) = \alpha(x,y)q(x,y) \tag{8.11}$$

We prove that $\pi(x)$ is the stationary distribution by showing it satisfies detailed balance. Let x and y be distinct states. (If $x = y$ it is trivial to verify the detailed balance equation.) So we must show $\pi(x)p(x,y) = \pi(y)p(y,x)$. This is immediate from (8.9).

Now consider the continuous case. So the state space is a subspace of R^d and $\pi(x)$ is a density with respect to Lebesgue measure on R^d. Note that the transition kernel is now a mix of a continuous and discrete measure. So trying to use detailed balance is problematic. We just verify the stationary equation. So let $A \subset R^d$. We must show

$$\int_A \pi(x)dx = \int K(x,A)\pi(x)dx \tag{8.12}$$

The right side is the sum of two terms - one is from when we accept the proposed new state and one from when we reject it. The acceptance term is

$$\int \left[\int_A \alpha(x,y)q(x,y)dy\right] \pi(x)dx \tag{8.13}$$

Given that we are in state x and that the proposed state is y, the probability of rejecting the proposed state is $1 - \alpha(x,y)$. So the probability we stay in x given that we are in x is

$$\int [1 - \alpha(x,y)]q(x,y)dy \tag{8.14}$$

So the rejection term is

$$\int_A \pi(x) \left[\int [1 - \alpha(x,y)]q(x,y)dy\right] dx \tag{8.15}$$

Note that the integral over x is only over A since when we reject we stay in the same state. So the only way to end up in A is to have started in A. Since $\int q(x,y)dy = 1$, the above equals

$$\int_A \pi(x)dx - \int_A \pi(x) \left[\int \alpha(x,y)]q(x,y)dy\right] dx \tag{8.16}$$

So we need to show

$$\int \left[\int_A \alpha(x,y) q(x,y) dy \right] \pi(x) dx = \int_A \pi(x) \left[\int \alpha(x,y)] q(x,y) dy \right] dx \tag{8.17}$$

In the right side we do a change of variables to interchange x and y. So we need to show

$$\int \left[\int_A \alpha(x,y) q(x,y) dy \right] \pi(x) dx = \int_A \pi(y) \left[\int \alpha(y,x)] q(y,x) dx \right] dy \tag{8.18}$$

If we integrate (8.9) over $x \in X$ and $y \in A$ we get the above. **QED**

Stop - Mon, 3/9

To apply our convergence theorem for Markov chains we need to know that the chain is irreducible and if the state space is continuous that it is Harris recurrent.

Consider the discrete case. We can assume that $\pi(x) > 0$ for all x. (Any states with $\pi(x) = 0$ can be deleted from the state space.) Given states x and y we need to show there are states $x = x_0, x_1, \cdots, x_{n-1}, x_n = y$ such that $\alpha(x_i, x_{i+1}) q(x_i, x_{i+1}) > 0$. If $q(x_i, x_{i+1}) > 0$ then $\alpha(x_i, x_{i+1}) > 0$. So it is enough to find states so that $q(x_i, x_{i+1}) > 0$. In other words we need to check that the transition function $q(x,y)$ is irreducible.

Now consider the continuous case. We want to show that the chain is π-irreducible. We start with a trivial observation. If $q(x,y) > 0$ for all x, y, then the chain is π-irreducible since for any x and any set A with $\int_A \pi(x) dx > 0$, the probability that if we start in x and reach A in just one step will be non-zero. This condition is too restrictive for many cases. Here is a more general sufficient condition.

Proposition 1 *Suppose that the state space S is connected in the following sense. Given $\delta > 0$ and $x, y \in S$ there exists states $y_0 = x, y_1, \cdots, y_{n-1}, y_n = y$ such that $|y_i - y_{i-1}| < \delta$ for $i = 1, 2, \cdots, n$ and the sets $B_\delta(y_i) \cap S$ have non-zero Lebesgue measure for $i = 0, 1, 2, \cdots, n$. Assume there is an $\epsilon > 0$ such that $|x - y| < \epsilon$ implies $q(x,y) > 0$. Then the Metropolis-Hasting chain is irreducible with respect to Lebesgue measure on S.*

Proof: Let $x_0 \in S$ and let $A \subset S$ have non-zero Lebesgue measure. Pick $y_n \in A$ such that the set $B_\delta(y_n) \cap A$ has non-zero Lebesgue measure. (This is possible since A has non-zero Lebesgue measure.) Let $y_1, y_2, \cdots, y_{n-1}$ be states as in the above sense of connectedness with

$\delta = \epsilon/3$. We will show $K^n(x_0, A) > 0$. We do this by only considering trajectories $x_0, x_1, x_2, \cdots, x_n$ such that $|x_i - y_i| < \delta$ for $i = 1, \cdots, n$. We further require $x_n \in A$. And finally we only consider trajectories for which all the proposed jumps were accepted. The probability of this set of trajectories is given by the integral of

$$q(x_0, x_1)\alpha(x_0, x_1)q(x_1, x_2)\alpha(x_1, x_2)\cdots q(x_{n-1}, x_n)\alpha(x_{n-1}, x_n) \tag{8.19}$$

where the region of integration is given by the constraints $|x_i - y_i| < \delta$ for $i = 1, 2, \cdots, n-1$ and $x_n \in B_\delta(y_n) \cap A$. Since $|y_{i-1} - y_i| < \delta$ the triangle inequality implies $|x_{i-1} - x_i| < 3\delta = \epsilon$. So we have $q(x_{i-1}, x_i) > 0$. Note that $q(x_{i-1}, x_i) > 0$ implies $\alpha(x_{i-1}, x_i) > 0$. So the integrand is strictly positive in the integral. The integral is over a set of non-zero Legesgue measure, so the integral is non-zero. **QED**

Finally we have

Proposition 2 *If the Metropolis-Harris chain is π-irreducible then it is Harris recurrent.*

A proof can be found in Robert and Casella. This is lemma 7.3 in their book in section 7.3.2.

We do not need to know the chain is aperiodic for our main use of it, but it is worth considering. If $U > \alpha(X_n, Y)$ then we stay in the same state. This happens with probability $1 - \alpha(X_n, Y)$. So as long as $\alpha(x, y) < 1$ on a set with non-zero probability (meaning what ???), the chain will be aperiodic. If $\alpha(x, y) = 1$ for all x, y, then $\pi(y)q(y, x) = \pi(x)q(x, y)$. But this just says that π satisfies detailed balance for the transition function q. So we would not be doing Metropolis-Hasting anyway. In this case we would need to study q to see if it was aperiodic.

Example (normal distribution): Want to generate samples of standard normal. Given $X_n = x$, the proposal distribution is the uniform distribution on $[x - 1, x + 1]$. So

$$q(x, y) = \begin{cases} \frac{1}{2} & \text{if } |x - y| \le 1, \\ 0, & \text{if } |x - y| > 1, \end{cases} \tag{8.20}$$

We have

$$\alpha(x, y) = \min\{\frac{\pi(y)}{\pi(x)}, 1\} = \frac{1}{2}\min\{\exp(-\frac{1}{2}y^2 + \frac{1}{2}x^2), 1\} \tag{8.21}$$

$$= \frac{1}{2}\begin{cases} \exp(-\frac{1}{2}y^2 + \frac{1}{2}x^2) & \text{if } |x| < |y|, \\ 1 & \text{if } |x| \ge |y| \end{cases} \tag{8.22}$$

Example (permutations): Consider permutations σ of $\{1, 2, \cdots, k\}$. A permutation is a bijective function on $\{1, 2, \cdots, k\}$. Instead of considering the uniform distribution on the set

of permutations as we did in an earlier example we consider a general probability measure. We write it in the form

$$\pi(\sigma) = \frac{1}{Z}\exp(w(\sigma)) \tag{8.23}$$

$w(\sigma)$ can be any function on permutation and can take on positive and negative values. An example of a possible w is the following. Let $s(\sigma)$ be the number of elements that are fixed by σ. Then let $w(\sigma) = \alpha s(\sigma)$. So depending on the sign of α we either favor or disfavor permutations that fix a lot of elements. The proposal distribution is to choose two distinct i, j uniformly and multiply the current permuation by the tranposition (i, j).

$$\alpha(\sigma, \sigma') = \min\{\frac{\pi(\sigma')}{\pi(\sigma)}, 1\} = \min\{\exp(w(\sigma') - w(\sigma)), 1\} \tag{8.24}$$

$$\tag{8.25}$$

Note that we only need to compute the change in $w()$. For "local" w this is a relatively cheap computation.

Example (Ising model): Fix a finite subset Λ of the lattice Z^d. At each site $i \in \Lambda$ there is a "spin" σ_i which takes on the values ± 1. The collection $\sigma = \{\sigma_i\}_{i\in\Lambda}$ is called a spin configuration and is a state for our system. The state space is $\{-1, 1\}^\Lambda$. The Hamiltonian $H(\sigma)$ is a function of configurations. The simplest H is the nearest neighbor H:

$$H(\sigma) = \sum_{<ij>} \sigma_i \sigma_j \tag{8.26}$$

We then define a probability measure on the spin configurations by

$$\pi(\sigma) = \frac{1}{Z}\exp(-\beta H(\sigma)) \tag{8.27}$$

The proposal distribution is defined as follows. We pick a site i uniformly from Λ. Then we flip the spin at i, i.e., we replace σ_i by $-\sigma_i$. So we only propose transitions between configurations that only only differ in one site. So $q(\sigma, \sigma') = 1/|\Lambda|$ when the spin configurations differ at exactly on site and it is zero otherwise. For two such configurations σ and σ' the acceptance probability is

$$\alpha(\sigma, \sigma') = \min\{\frac{\pi(\sigma')}{\pi(\sigma)}, 1\} = \min\{\exp(-\beta[H(\sigma') - H(\sigma)], 1\} \tag{8.28}$$

Note that there is lots on cancellation in the differece of the two Hamiltonians. This computation takes a time that does not depend on the size of Λ.

Example: QFT

8.3 The independence sampler

The independence sampler is a special case of the Metropolis-Hasting algorithm. In the independence sampler the proposal distribution does not depend on x, i.e., $q(x,y) = g(y)$. So the acceptance probability becomes

$$\alpha(x,y) = \min\{\frac{\pi(y)g(x)}{\pi(x)g(y)}, 1\} \tag{8.29}$$

Suppose that there is a constant C such that $\pi(x) \leq Cg(x)$. In this setting we could do the acceptance-rejection algorithm. It will generate independent samples of $\pi(x)$ and the acceptance rate will be $1/C$. By contrast the independence sampler will generate dependent samples of $\pi(x)$.

Stop - Wed, 3/9

Proposition 3 *Consider the independence sampler with proposal distribution $g(x)$ and stationary distribution $\pi(x)$. Suppose there is a constant C such that $\pi(x) \leq Cg(x)$ for all $x \in S$. Let $\pi_0(x)$ be any initial distibution and let $\pi_n(x)$ be the distribution at time n. Then*

$$||\pi_n - \pi||_{TV} \leq 2(1 - \frac{1}{C})^n \tag{8.30}$$

Proof: We will only consider the case that the initial distribution is absolutely continuous with respect to Lebesgue measure. So the initial distribution is $\pi(x)dx$. Note that in this case the subsequent distributions π_n will be absolutely continuous with respect to Lebesgue measure. **Explain this.**

For convenience let $\epsilon = \frac{1}{C}$. So our bound can be rewritten as $g(x) \geq \epsilon\pi(x)$. Since $q(x,y) = g(y)$, we have

$$\begin{aligned} \alpha(x,y)q(x,y) &= \min\{\frac{\pi(y)g(x)}{\pi(x)g(y)}, 1\}g(y) = \min\{\frac{\pi(y)g(x)}{\pi(x)}, g(y)\} & (8.31) \\ &\geq \min\{\frac{\pi(y)\epsilon\pi(x)}{\pi(x)}, \epsilon\pi(y)\} = \epsilon\pi(y) & (8.32) \end{aligned}$$

Let $\rho(x)$ be a probability density. The transition kernel takes it to another density, and we will denote this new density by $K\rho$. So K is a linear operator on integrable functions on R^d.

Let P be the linear operator which maps a probability density $\rho(x)$ to the probability density $\pi(x)$. So P is a projection. For a general integrable function

$$(P\rho)(x) = \pi(x) \int_X \rho(y) dy \tag{8.33}$$

Our previous bound shows that for any probability density ρ, $K\rho - \epsilon P\rho$ is a non-negative function. Its integral is $1 - \epsilon$. So if we define another linear operator by

$$R = \frac{1}{1-\epsilon}[K - \epsilon P] \tag{8.34}$$

then R will map a probability density into another probability density. Note that $K = \epsilon P + (1-\epsilon)R$. A straightforward induction argument shows that

$$K^n = \sum_{k=1}^{n} K^{n-k} P (1-\epsilon)^{k-1} R^{k-1} + (1-\epsilon)^n R^n \tag{8.35}$$

Note that $P\rho = \pi$ for any probability distribution ρ, and $K\pi = \pi$. So $K^j P\rho = \pi$ for any j. So for any initial distribution π_0,

$$\pi_n = K^n \pi_0 = \pi \sum_{k=1}^{n} (1-\epsilon)^{k-1} + (1-\epsilon)^n R^n \pi_0 = [1 - (1-\epsilon)^n]\pi + (1-\epsilon)^n R^n \pi_0 \tag{8.36}$$

So

$$\pi_n - \pi = -(1-\epsilon)^n \pi + (1-\epsilon)^n R^n \pi_0 \tag{8.37}$$

Since $R^n \pi_0$ is a probability density, the L^1 norm of the above is bounded by $2(1-\epsilon)^n$. **QED**

8.4 The Gibbs sampler

In this section change notation and let $f(x)$ denote the distribution we want to sample from, in place of our previous notation $\pi(x)$.

We first consider the two-stage Gibbs sampler. We assume that the elements of the state space are of the form (x, y). The probability distribution we want to sample from is $f(x, y)$. Recall that the marginal distributions of X and Y are given by

$$f_X(x) = \int f(x,y)\, dy, \quad f_Y(y) = \int f(x,y)\, dx \tag{8.38}$$

and the conditional distributions of X given Y and Y given X are

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}, \quad f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \tag{8.39}$$

Note that if we only know $f(x,y)$ up to an overall constant, we can still compute $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$.

The two-stage Gibbs sampler *Given that we are in state (X_n, Y_n), we first generate Y_{n+1} from the distribution $f_{Y|X}(\cdot|X_n)$. Then we generate X_{n+1} from the distribution $f_{X|Y}(\cdot|Y_{n+1})$. These two stages make up one time step for the Markov chain. So the transition kernel is*

$$K(x,y;x',y') = f_{Y|X}(y'|x) f_{X|Y}(x'|y') \tag{8.40}$$

Proposition 4 *$f(x,y)$ is the stationary distribution of the two-stage Gibbs sampler.*

Proof: We just show that $Kf = f$.

$$\begin{aligned}
(Kf)(x',y') &= \int\int f(x,y) K(x,y;x',y') dxdy && (8.41)\\
&= \int\int f(x,y) f_{Y|X}(y'|x) f_{X|Y}(x'|y') dxdy && (8.42)\\
&= \int\int f(x,y) \frac{f(x,y')}{f_X(x)} \frac{f(x',y')}{f_Y(y')} dxdy && (8.43)\\
&= \int f_X(x) \frac{f(x,y')}{f_X(x)} \frac{f(x',y')}{f_Y(y')} dx && (8.44)\\
&= \int f(x,y') \frac{f(x',y')}{f_Y(y')} dx && (8.45)\\
&= f(x',y') && (8.46)
\end{aligned}$$

QED

Remark: The two-stage Gibbs sampler does not satisfy detailed balance in general.

Example: (Bivariate normal) We consider the bivariate normal (X,Y) with joint density

$$f(x,y) = c \exp(-\frac{1}{2}x^2 - \frac{1}{2}y^2 - \alpha xy) \tag{8.47}$$

where α is a parameter related to the correlation of X and Y. Argue that

$$f_{Y|X}(y|x) = c(x) \exp(-\frac{1}{2}(y+\alpha x)^2) \tag{8.48}$$

So for the first stage in the Gibbs sampler, we generate Y_{n+1} from a standard normal distribution with mean $-\alpha X_n$. We have

$$f_{X|Y}(x|y) = c(y) \exp(-\frac{1}{2}(x+\alpha y)^2) \tag{8.49}$$

So for the second stage, we generate X_{n+1} from a standard normal distribution with mean $-\alpha Y_{n+1}$.

We now consider the multi-stage Gibbs sampler. Now suppose that the points in the state space are of the form $x = (x_1, x_2, \cdots, x_d)$. We need to consider the conditional distribuion of X_i given all the other X_j. To keep the notation under control we will write

$$f_{X_i|X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_d}(x_i|x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_d) = f_i(x_i|x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_d) \quad (8.50)$$

Again we emphasize that we can compute these conditional distributions even if we only know $f(x_1,\cdots,x_d)$ up to an unknown constant.

Multi-stage Gibbs sampler: *The algorithm has d stages and proceeds as follows.*
(1) Given $(X_1^n,\cdots,X_d^n)$ *we sample* X_1^{n+1} *from* $f_1(\cdot|X_2^n,\cdots,X_d^n)$.
(2) Then we sample X_2^{n+1} *from* $f_2(\cdot|X_1^{n+1},X_3^n,\cdots,X_d^n)$.
(j) Continuing we sample X_j^{n+1} *from* $f_j(\cdot|X_1^{n+1},\cdot,X_{j-1}^{n+1},X_{j+1}^n,\cdots,X_d^n$.
(p) In the last step we sample X_d^{n+1} *from* $f_d(\cdot|X_1^{n+1},\cdots X_{d-1}^{n+1})$.

Before we show that the stationary distribution of this algorithm is f, we consider some variations of the algorithm. Let K_j be the transition kernel corresponding to the jth step of the multi-stage Gibbs sampler. So

$$K_j(x_1,x_2,\cdots,x_p;x_1',x_2',\cdots,x_d') = f_j(x_j'|x_1,x_2,\cdots,x_{j-1},x_{j+1},\cdots,x_p)\prod_{i=1:i\neq j}^{p}\delta(x_i-x_i') \quad (8.51)$$

If we think of K_j as a linear operator, then the multi-stage Gibbs sampler is $K_dK_{d-1}\cdots K_2K_1$.

Here is another Gibbs sampler which for lack of a standard name we will call the randomized Gibbs sampler. Fix some probability distribution p_i on $\{1,2,\cdots,d\}$. Given that we are in state $(X_1^n,\cdots,X_d^n)$, we first pick $i\in\{1,2,\cdots,d\}$ according to this distribution. Then we sample X_i^{n+1} from $f_i(\cdot|X_1^n,\cdot,X_{i-1}^n,X_{i+1}^n,\cdots,X_d^n)$. For $l\neq i$, $X_l^{n+1}=X_l^n$. The transition kernel for this algorithm is

$$K = \sum_{i=1}^{d} p_iK_i \quad (8.52)$$

Proposition 5 $f(x_1,x_2,\cdots,x_d)$ *is the stationary distribution of the multi-stage Gibbs sampler and of the randomized Gibbs sample for any choice of the distribution* p_i.

Stop - Mon, 3/21

Proof: We only need to show that for all j, $K_j f = f$. So we compute:

$$
\begin{aligned}
& (K_j f)(x'_1, x'_2, \cdots, x'_d) && (8.53)\\
= & \int \cdots \int f(x_1, x_2, \cdots, x_d) K_j(x_1, x_2, \cdots, x_d; x'_1, x'_2, \cdots, x'_d) dx_1 dx_2 \cdots dx_d && (8.54)\\
= & \int \cdots \int f(x_1, x_2, \cdots, x_d) f_j(x'_j | x_1, x_2, \cdots, x_{j-1}, x_{j+1}, \cdots, x_d) \prod_{i=1:i\neq j}^{d} \delta(x_i - x'_i) && (8.55)\\
= & f_j(x'_j | x'_1, x'_2, \cdots, x'_{j-1}, x'_{j+1}, \cdots, x'_d) \int f(x'_1, x'_2, \cdots, x'_{j-1}, x_j, x'_{j+1}, x'_d) dx_j && (8.56)\\
= & f(x'_1, x'_2, \cdots, x'_d) && (8.57)
\end{aligned}
$$

Note that the random stage Gibbs sampler has f as the stationary distrbution for any choice of the p_i. We need to take all $p_i > 0$ to have any chance that the chain is irreducible. The simplest choice for the p_i is to use the uniform distribution on $\{1, 2, \cdots, d\}$. Why would we do anything else? In the following example we will see a case where we might want to use a non-uniform distribution.

Caution: There are lots of variations on the Gibbs sampler, but one should be careful. Here is one that does not work.

WRONG two-stage Gibbs sampler: Given that we are in state (X_n, Y_n), we first generate Y_{n+1} from the distribution $f_{Y|X}(\cdot|X_n)$. Then we generate X_{n+1} from the distribution $f_{X|Y}(\cdot|Y_n)$. These two stages make up one time step for the Markov chain. So the transition kernel is

$$K(x, y; x', y') = f_{Y|X}(y'|x) f_{X|Y}(x'|y) \tag{8.58}$$

Note that the difference with the correct two-stage Gibbs sampler is that we we generate X_{n+1} from $f_{X|Y}(\cdot|Y_n)$ rather than $f_{X|Y}(\cdot|Y_{n+1})$.

Here is an example to illustrate how the above algorithm is wrong. Take $f(x, y)$ to be the uniform distribution on the three points $(0, 0), (0, 1), (1, 0)$. Explain this.

Remark: The d-stage Gibbs sampler requires that the states have the structure $(x_1, x_2, \cdots, x_d)$. However this does mean that the state space has to be a subset of R^d. Some of the x_i could be vectors or even something stranger.

Example: We consider the Ising model that we considered in a previous example. The integer d is not the number of dimensions. It is the number of sites in Λ. For $j \in \Lambda$, f_j is the

conditional distribution of σ_j given the values of all the other spins. We compute this in the usual way (joint density over marginal) to get

$$f_j(\sigma_j|\sigma_{\Lambda\setminus j}) = \frac{\exp(-\beta H(\sigma))}{\sum_{s_j}\exp(-\beta H(\hat{\sigma}))} \tag{8.59}$$

where s_j is summed over just $-1, 1$ and $\hat{\sigma}$ equals σ_i for all sites $i \neq j$ and equals s_j at site j. The algorithm applies to any H, but there are some nice cancellations if H is "local." We illustrate this by considering the nearest neighbor H. Any term in H that does not involve site j cancels in the numerator and the denominator. The result is just

$$f_j(\sigma_j|\sigma_{\Lambda\setminus j}) = \frac{\exp(-\beta\sigma_j \sum_{k:|k-j|=1}\sigma_k)}{\exp(-\beta\sum_{k:|k-j|=1}\sigma_k) + \exp(\beta\sum_{k:|k-j|=1}\sigma_k)} \tag{8.60}$$

So computing f_j takes a time that is $O(1)$, independent of the size of Λ. But just how fast the algorithm mixes depends very much on the size of Λ and on β. For the multi-stage algorithm each time steps take a time of order $|\Lambda|$. For the random-stage algorithm each time step only takes a time $O(1)$, but it will take $O(|\Lambda|)$ times steps before we have changed a significant fraction of the spins.

Now suppose we want to compute the expected value of $F(\sigma)$ in the Ising model and $F(\sigma)$ only depends on a few spins near the center of Λ. Then we may want to choose the distribution p_i so that the sites near the center have higher probability than the sites that are not near the center.

Remark: As the example above shows, d is not always the "dimension" of the model.

Example: (loosely based on example 6.6 in Rubenstein and Kroese, p. 177) For $i = 1, 2, \cdots, d$, let $p_i(x_i)$ be a discrete probability function on the non-negative integers. If $X_1, X_2, \cdots, X_d$ were independent with these distributions, then the joint distribution would be just the product of the $p_i(x_i)$. This is trivial to simulate. We are interested in something else. Fix a positive integer m. We restrict the sample sample to the d-tuples of non-negative integers $x_1, x_2, \cdots, x_d$ such that $\sum_{i=1}^d x_i = m$. We can think of this as the conditional distribution of $X_1, \cdots, X_d$ given that $\sum_i X_i = m$. So we want to simulate the joint pdf given by

$$f(x_1, \cdots, x_d) = \frac{1}{Z}\prod_{i=1}^d p_i(x_i) \tag{8.61}$$

when $\sum x_i = m$ and $f() = 0$ otherwise. The constant Z is defined by ... Since $X_d = m - \sum_{i=1}^{d-1} X_i$, we can work with just $X_1, X_2, \cdots, X_{d-1}$. Their joint distribution is

$$f(x_1, \cdots, x_{d-1}) = \frac{1}{Z}\, p_d(m - \sum_{i=1}^{d-1} x_i)\, \prod_{i=1}^{d-1} p_i(x_i) \tag{8.62}$$

for $x_1, \cdots, x_{d-1}$ whose sum is less than or equal to m. Then their sum is greater than m, $f(x_1, \cdots, x_{d-1}) = 0$. All we need to run the Gibbs sampler are the conditional distributions of X_j given the other X_i. They are given by

$$f_j(x_j|x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_{d-1}) \propto f_j(x_j) f_d(m - \sum_{i=1}^{d-1} x_i) \tag{8.63}$$

If we let $m' = m - \sum_{i=1:i\neq j}^{d-1} x_i$, then the right side is equal to $f_j(x_j) f_d(m' - x_j)$. We need to compute the constant to normalize this, but that takes only a single sum on x_j. (And for some f_j, f_d can be done explicitly.)

Irreducibility: brief discussion of irreducibility for the Gibbs sample. Gap in notes here.

8.5 Slice sampler

The slice sampler is in some sense a special case of the Gibbs sampler. Suppose we want to sample from $f(x)$ where x ranges over X. We consider a new distribution: the new state space is a subspace of $X \times R$, namely,

$$S = \{(x, u) : 0 \leq u \leq f(x)\} \tag{8.64}$$

This can be thought of as the area under the graph of f. The new distribution is the uniform measure on S. The key observation is that with this distribution on (X, U), the marginal distribution of X is $f(x)$. So if we can construct a Markov chain (X_n, U_n) with the uniform measure on S as its stationary measure then we can just look at X_n and long time averages of random variables on X will converge to their expectation with respect to $f(x)$. We use the two-stage Gibbs sampler. So we need the conditional distributions $f_{U|X}(u|x)$ and $f_{X|U}(x|u)$. They are both uniform. More precisely, the distribution of U given $X = x$ is uniform on $[0, f(x)]$, and the distribution of X given $U = u$ is uniform on $\{x : u \leq f(x)\}$.

Slice sampler (single slice) *Given that we are in state (X_n, U_n), we first generate U_{n+1} from the uniform distribution on $[0, f(X_n)]$. Then we generate X_{n+1} from the uniform distribution on $\{x : U_{n+1} \leq f(x)\}$.*

Remark: Suppose that the density we wish to simulate is given by $cf(x)$ where c is an unknown constant. We can still take

$$S = \{(x, u) : 0 \leq u \leq f(x)\} \tag{8.65}$$

and put the uniform distribution on S. The density function is $\frac{1}{c} 1_S$. The constant c is unknown, but that will not matter. The marginal density of X is still $f(x)$.

Example: Let $f(x) = ce^{-x^2/2}$, the standard normal. Given (X_n, U_n) it is trivial to sample U_{n+1} uniformly from $[0, f(X_n)]$. Next we need to sample X_{n+1} uniformly from $\{x : U_{n+1} \leq f(x)\}$. This set is just the interval $[-a, a]$ where a is given by $U_{n+1} = f(a)$. We can trivially solve for a.

The first step of the slice sampler, generating U_{n+1}, is always easy. The second step, generating X_{n+1}, may not be feasible at all since the set $\{x : U_{n+1} \leq f(x)\}$ may be very complicated. For example suppose we want to sample

$$f(x) = c\frac{1}{1+x^2}\exp(-x^2/2) \tag{8.66}$$

The set will be an interval, but finding the endpoints requires solving an equation like $\exp(-x^2/2)/(1+x^2) = u$. This could be done numerically, but the set could be even more complicated. There is a generalization that may work even when this second step is not feasible for the single slice sampler.

Assume that $f(x)$ can be written in the form

$$f(x) = \prod_{i=1}^{d} f_i(x) \tag{8.67}$$

where the $f_i(x)$ are non-negative but need not be probability densities. We then introduce a new random variable (sometimes called auxillary variables) for each f_i. So the new state space is a subspace of $X \times R^d$ and is given by

$$S = \{(x, u_1, u_2, \cdots, u_d) : 0 \leq u_i \leq f_i(x),\ i = 1, 2, \cdots, d\} \tag{8.68}$$

We use the uniform distribution on S. The key observation is that if we integrate out $u_1, u_2, \cdots, u_d$, we just get $f(x)$. So the marginal distribution of X will be $f(x)$. For the Markov chain we use the $d+1$ dimensional Gibbs sample.

Example: Let

$$f(x) = c\frac{1}{1+x^2}\exp(-x^2/2) \tag{8.69}$$

Let

$$f_1(x) = \frac{1}{1+x^2}, \quad f_2(x) = \exp(-x^2/2) \tag{8.70}$$

Note that we are dropping the c. We sample U_1^{n+1} uniformly from $[0, f_1(X^n)]$. Then we sample U_2^{n+1} uniformly from $[0, f_2(X^n)]$. Finally we need to sample X^{n+1} uniformly from

$$\{x : U_1^{n+1} \leq f_1(x),\ U_2^{n+1} \leq f_2(x)\} \tag{8.71}$$

This set is just an interval with endpoints that are easily computed.

Stop - Wed, March 23

The slice sampler can be used when the initial distribution $f(x)$ is discrete as the next example shows.

Example: Consider the density $f(x) = c\exp(-\alpha x^2)$ where $\alpha > 0$ and $x = 0, 1, 2, \cdots$. Note that the constant c cannot be computed analytically. We would have to compute it numerically. For the slice sampler we can just drop c. The first stage is to generate U_{n+1} uniformly from $[0, f(X_n)]$. Then we generate X_{n+1} uniformly from the set of non-negative integers k such that $U_{n+1} \leq f(k)$.

8.6 Bayesian statistics and MCMC

We start with a triviality which is often called Bayes rule. Given two random variables (which can be random vectors), we have

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \quad f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \tag{8.72}$$

So

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)} \tag{8.73}$$

This is often written as

$$f_{Y|X}(y|x) \propto f_{X|Y}(x|y)f_Y(y) \tag{8.74}$$

with the understanding that the constant of proportionality depends on x. In Bayesian statistics it is often written in the more abbreviated form

$$f(y|x) \propto f(x|y)f(y) \tag{8.75}$$

"This particular style of notation is typical in Bayesian analysis and can be of great descriptive value, despite its apparent ambiguity" - Rubinstein and Kroese.

Now suppose we have a probability distribution for x, which is typically a vector, that depends on some parameters $\theta = (\theta_1, \cdots, \theta_d)$. Often the vector x is a sample $x_1, x_2, \cdots, x_n$ that comes from performing some experiment n times. We don't know θ. In statistics we want to use the value of x that results from our experiment to estimate the unknown parameters θ. The Bayesian statistician puts a probability distribution on θ, $f(\theta)$, that is supposed to encode all the information we have about how likely we think different values of θ are **before** we do the experiment. $f(\theta)$ is called the *prior distribution*. Now we do the experiment, and so we have a particular value for x. We want to replace the prior distribution on θ by a distribution that incorporates the knowledge of x. The natural distribution is $f(\theta|x)$. This is called the *posterior* distribution of θ. By Bayes rule

$$f(\theta|x) \propto f(x|\theta)f(\theta) \tag{8.76}$$

where the constant of proportionality depends on x. The conditional density $f(x|\theta)$ is called the *likelihood*. We typically know this function quite explicitly. For example, if $f(x|\theta)$ comes from independent repetitions of the same experiment, then

$$f(x|\theta) = f(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta) \tag{8.77}$$

where $f_X(x|\theta)$ is the distribution of X for one performance of the experiment. So Bayes rule says

$$f(\theta|x) \propto \left[\prod_{i=1}^{n} f_X(x_i|\theta)\right] f(\theta) \tag{8.78}$$

Given the data x this gives the joint distribution of the parameters $\theta_1, \cdots, \theta_d$. To run a Gibbs sampler we need the conditional distribution of each θ_i given the other θ_j, $j \neq i$. The constant of proportionality in Bayes rule is often impossible to compute analytically, but this does not matter for the Gibbs sampler.

Example : We have a coin with probability θ of getting heads. However, we do not know θ. We flip it n times, let $X_1, X_2, \cdots, X_n$ be 1 for heads, 0 for tails. If we are given a value for θ, then the distribution of $X_1, X_2, \cdots, X_n$ is just

$$f(x|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^s(1-\theta)^{n-s} \tag{8.79}$$

where x is short for $x_1, x_2, \cdots, x_n$ and s is defined to be $\sum_{i=1}^{n} x_i$. If we have no idea what θ is, a reasonable choice for the prior distribution for θ is to make it uniform on $[0, 1]$. Now suppose we flip the coin n times and use the resulting "data" $x_1, \cdots, x_n$ to find a better distribution for θ that incorporates this new information, i.e., find the posterior distribution. The posterior is given by

$$f(\theta|x) \propto f(x|\theta)f(\theta) = \theta^s(1-\theta)^{n-s}1_{[0,1]}(\theta) \tag{8.80}$$

where $s = \sum_{i=1}^n X_i$. If n is large then s will be large too and this density will be sharply peaked around s/n.

In this example above the formula for the posterior is quite simple and in particular it is trivial to compute the normalizing constant. In many actual applications this is not the case. Often θ is multidimensional and so just computing the normalizing constant requires doing a multidimensional integral which may not be tractable. We still want to be able to generate samples from the posterior. For example we might want to compute the mean of θ from the posterior and maybe find confidence interval for it. We can try to use MCMC, in particular the Gibbs sampler, to do this.

Example: This is similar to the coin example above but with more parameters. We have a die with probabilities $\theta_1, \theta_2, \cdots, \theta_6$ of getting $1, 2, \cdots, 6$. So the sum of the θ_i must be 1. We role the die n times and let $x_1, x_2, \cdots, x_n$ be the numbers we get. So the x_i take values in $\{1, 2, 3, 4, 5, 6\}$. Putting a prior distribution on the θ_i is a little tricky since we have the constraint that they must sum to 1. Here is one approach. We would like to assume the die is close to being fair and we have no prior reason to think that a particular number is more likely than any other number. Take $\phi_1, \cdots, \phi_6$ to be independent and identically distributed with distribution $g(\phi)$ where ϕ is peaked around $1/6$. So the joint distribution of the ϕ_i is $\prod_i g(\phi_i)$. Then we just set $\theta_i = \phi_i / \sum_j \phi_j$. We now think of the ϕ_i as the parameters.

We have

$$f(x|\theta) = \prod_{i=1}^n \theta_{x_i} = \prod_{j=1}^6 \theta_j^{n_j} \tag{8.81}$$

and so

$$f(x|\phi) = [\sum_{j=1}^6 \phi_j]^{-n} \prod_{j=1}^6 \phi_j^{n_j} \tag{8.82}$$

where n_j is the number of x_i equal to j. We have used the fact that $\sum_{j=1}^6 n_j = n$. So Bayes rule says

$$f(\phi_1, \cdots, \phi_6|x) \propto [\sum_{j=1}^6 \phi_j]^{-n} \prod_{j=1}^6 [\phi_j^{n_j} g(\phi_j)] \tag{8.83}$$

We would like to compute things like the expected value of each θ_i. This would give us an idea of how unfair the die is and just how it is "loaded". We do this by generating samples of $(\phi_1, \cdots, \phi_6)$. We can use the Gibbs sampler. We need the conditional distribution of each ϕ_i given the other ϕ. Up to a normalization constant this is

$$[\phi_i + \Phi]^{-n} \phi_i^{n_i} g(\phi_i) \tag{8.84}$$

where $\Phi = \sum_{j \neq i} \phi_j$.

Example: Zero-inflated poisson process - handbook p. 235.

Review Poisson proccesses, Poisson RV's, and gamma distribution

$Gamma(w, \lambda)$ has pdf

$$f(x) = \frac{\lambda^w}{\Gamma(w)} x^{w-1} e^{-\lambda x} \tag{8.85}$$

Hierarchical models: Suppose we have a parameter λ which we take to be random. For example it could have a $Gamma(\alpha, \beta)$ distribution. Now we go one step further and make β random, say with a $Gamma(\gamma, \delta)$ distribution. So

$$f(\lambda|\beta) = Gamma(\alpha, \beta), \tag{8.86}$$
$$f(\beta) = Gamma(\gamma, \delta) \tag{8.87}$$

and so the prior is

$$f(\lambda, \beta) = f(\lambda|\beta) f(\beta) = \cdots \tag{8.88}$$

Example The following example appears in so many books and articles it is ridiculuous. But it is still a nice example. A nuclear power plant has 10 pumps that can fail. The data consists of an observation time t_i and the number of failures x_i for each pump that have occured by time t_i.

A natural model for the times at which a single pump fail is a Poisson process with parameter λ. We only observe the process at a single time, and the number of failures that have occured by that time is a Possion random variable with parameter λt_i. One model would be to assume that all the pumps have the same failure rate, i.e., the same λ. This is an unrealistic assumption. Instead we assume that each pump has its own failure rate λ_i. The λ_i are assumed to be random and independent, but with a common distribution. We take this common distribution to be $Gamma(\alpha, \beta)$ where α is a fixed value but β is random with distribution $Gamma(\gamma, \delta)$. γ and δ are numbers. The parameters θ here are $\lambda_1, \cdots, \lambda_{10}, \beta$. From now on we write $\lambda_1, \cdots, \lambda_{10}$ as λ. Note that

$$f(\lambda, \beta) = f(\lambda|\beta) f(\beta) \tag{8.89}$$

and

$$f(x|\lambda, \beta) = f(x|\lambda) \tag{8.90}$$

pump	1	2	3	4	5	6	7	8	9	10
Number failures	5	1	5	14	3	19	1	1	4	22
observation time	94.32	15.72	62.88	125.76	5.24	31.44	1.05	1.05	2.10	10.48

Table 8.1:

and we have

$$f(x|\lambda) = \prod_{i=1}^{10} \left[\frac{(\lambda_i t_i)^{x_i}}{x_i!} e^{-\lambda_i t_i} \right] \tag{8.91}$$

Bayes rule says

$$\begin{aligned} f(\lambda, \beta|x) &\propto f(x|\lambda, \beta) f(\lambda, \beta) && (8.92)\\ &= f(x|\lambda, \beta) f(\lambda|\beta) f(\beta) && (8.93)\\ &= \prod_{i=1}^{10} \left[\frac{(\lambda_i t_i)^{x_i}}{x_i!} e^{-\lambda_i t_i} \right] \prod_{i=1}^{10} Gamma(\lambda_i|\alpha, \beta) Gamma(\beta|\gamma, \delta) && (8.94)\\ &= \prod_{i=1}^{10} \left[\frac{(\lambda_i t_i)^{x_i}}{x_i!} e^{-\lambda_i t_i} \right] \prod_{i=1}^{10} \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_i^{\alpha-1} e^{-\lambda_i \beta} \right] \frac{\delta^\gamma}{\Gamma(\gamma)} \beta^{\gamma-1} e^{-\delta\beta} && (8.95)\\ &\propto \prod_{i=1}^{10} \left[\lambda_i^{x_i+\alpha-1} e^{-\lambda_i(t_i+\beta)} \right] \beta^{10\alpha+\gamma-1} e^{-\delta\beta} && (8.96) \end{aligned}$$

where the constant of proportionality depends on the x_i and t_i and on the constants γ, δ.

We want to compute the posterior distribution of the parameters. We are particularly interested in the mean of the distribution of the λ_i, i.e., the mean of $Gamma(\alpha, \beta)$. The mean of this gamma distribution with fixed α, β is α/β. So we need to compute the mean of α/β over the posterior distribution. We can write this as a ratio of high dimensional (ten or eleven) integrals, but that is hard to compute. So we use the Gibbs sampler to sample λ, β from the posterior. Note that this is an 11 dimensional sampler. So we need the conditional distributions of each λ_i and of β.

$$\lambda_i|\beta, t_i, x_i \sim Gamma(x_i + \alpha, t_i + \beta), \tag{8.97}$$

$$\beta|\lambda \sim Gamma(\gamma + 10\alpha, \delta + \sum_{i=1}^{10} \lambda_i) \tag{8.98}$$