MTH 7241 Fall 2022: Prof. C. King

Notes 8: Markov Chain Monte Carlo

0.1 The basic idea

Persi Diaconis titled his review article 'The Markov Chain Monte Carlo Revolution', referring to the fact that MCMC techniques have become pervasive and essential in most aspects of Data Analysis and Applied Probability. By using powerful computing resources the MCMC technique allows answers to be found for highly complicated problems involving statistical optimization.

We will describe the basic MCMC algorithm for discrete Markov chains on finite state spaces, and provide several concrete examples. Let Ω be a finite state space, and let π be a probability distribution on Ω. The goal is to generate random samples from π, meaning to generate elements of Ω which are chosen randomly with probability distribution π. The MCMC method proceeds by constructing an ergodic Markov chain on Ω for which π is the stationary distribution. By running the Markov chain we generate a sequence of states $X_1, X_2, \dots$ whose distribution is guaranteed to converge to π. So for any $\epsilon > 0$, there is N sufficiently large such that the sequence $X_N, X_{N+1}, \dots$ has distribution close to π, in the sense that

$$\sum_{j \in \Omega} |\mathbb{P}(X_{N+k} = j) - \pi(j)| \leq \epsilon$$

for all $k \geq 0$.

0.2 The basic algorithm

We assume for definiteness that $\pi(j) > 0$ for all $j \in \Omega$. We also assume that $Q = (q_{ij})$ is a known symmetric irreducible aperiodic transition matrix on Ω. Being symmetric, the stationary distribution of Q is uniform on Ω. The MCMC algorithm will 'transform' Q into a new Markov chain with the desired stationary distribution π. The algorithm constructs a sequence $X_0, X_1, X_2, \dots$ where X_0 is any chosen initial state, and X_{k+1} is constructed from X_k using the following steps:
1) randomly select $Y \in \Omega$ according to the probability distribution Q:

$$\mathbb{P}(Y = j \,|\, X_k = i) = q_{ij}$$

2) define the acceptance probability

$$\begin{aligned} \alpha &= \min\left\{1, \frac{\pi(Y)}{\pi(X_k)}\right\} \\ &= \min\left\{1, \frac{\pi(j)}{\pi(i)}\right\} \quad \text{if } Y = j,\ X_k = i \end{aligned}$$

3) define

$$X_{k+1} = \begin{cases} Y & \text{with probability } \alpha \\ \\ X_k & \text{with probability } 1 - \alpha \end{cases}$$

(more concretely, we flip a weighted coin with probability of Heads equal to α; if it comes up Heads we set $X_{k+1} = Y$, if Tails we set $X_{k+1} = X_k$).

This protocol defines the off-diagonal entries of the transition matrix $P = (p_{ij})$ for the new Markov chain, namely

$$p_{ij} = \mathbb{P}(X_{k+1} = j \,|\, X_k = i) = q_{ij} \, \min\left\{1, \frac{\pi(j)}{\pi(i)}\right\} \quad \text{for } i \neq j$$

We then define the diagonal entries in order to make the matrix stochastic:

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij}$$

Theorem 1 *For all $i, j \in \Omega$,*

$$\pi(i) p_{ij} = \pi(j) p_{ji}$$

Proof: It is sufficient to assume that $i \neq j$, in which case

$$\begin{aligned} \pi(i) p_{ij} &= \pi(i) q_{ij} \, \min\left\{1, \frac{\pi(j)}{\pi(i)}\right\} \\ &= q_{ij} \min\left\{\pi(i), \pi(j)\right\} \\ &= q_{ji} \pi(j) \, \min\left\{1, \frac{\pi(i)}{\pi(j)}\right\} \quad \text{(since } Q \text{ is symmetric)} \\ &= \pi(j) p_{ji} \end{aligned}$$

$\boxed{\text{QED}}$

Remarks: Theorem 1 shows that the new Markov chain is reversible, and therefore it follows that π is its stationary distribution. So $\{X_k\}$ is the promised Markov chain that converges to π. Note also that the computation of p_{ij} requires only knowledge of the ratio $\pi(j)/\pi(i)$. We will see that there are many applications where these ratios can be easily computed, while the probabilities $\pi(i)$ are themselves difficult to find, because the overall normalization is hard to compute. One important case is Bayesian inference.

Example 1: sampling from a multinomial distribution

Suppose that X is discrete with $\mathrm{Ran}(X) = \{1, 2, \ldots, K\}$, and pdf

$$\mathbb{P}(X = k) = q_k, \qquad k = 1, \ldots, K.$$

We will use MCMC to estimate the expected value $\mathbb{E}[X^2]$. So our stationary pdf for the Markov chain is $\pi = (q_1, \ldots, q_K)$. We choose the reference transition to be uniform on the state space, so

$$q_{ij} = \frac{1}{K} \qquad \text{for all } i, j = 1, \ldots, K$$

Then the transition probability for MCMC is

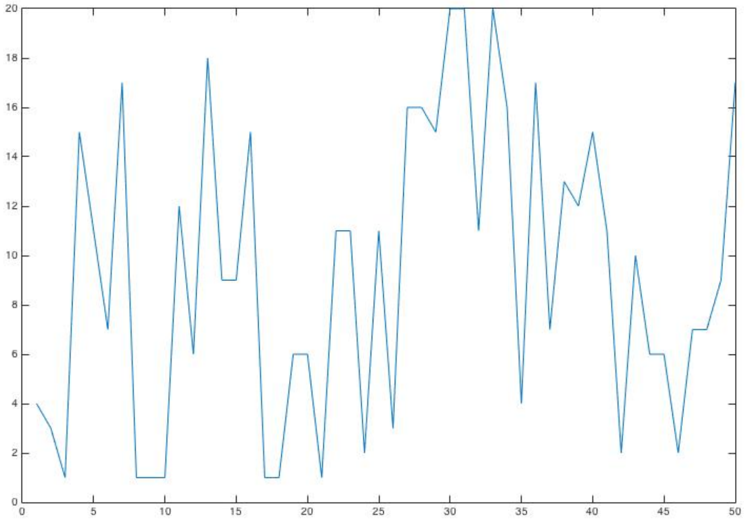
$$p_{ij} = \begin{cases} \frac{1}{K} \min\left(1, \frac{q_j}{q_i}\right) & \text{for all } i \neq j \\ 1 - \sum_{k \neq i} p_{ik} & \text{for } j = i \end{cases}$$

So starting with an initial state X_0, we simulate a sequence $X_1, X_2, \ldots$ using this transition matrix.

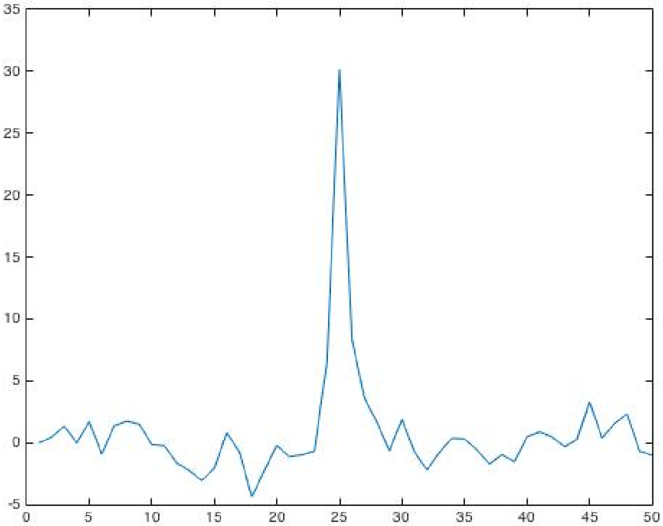
We then estimate the mixing time of the chain – this is the number of steps s such that X_n and X_{n+s} are approximately independent. Then we estimate the expected value as

$$\mathbb{E}[X^2] \simeq \frac{1}{N}\sum_{n=1}^{N} X_{ns}^2$$

Let's see a concrete example. Take $K = 20$, and generate a random distribution $(q_1, \ldots, q_{20})$. Choose initial state $X_0 = 4$, and generate a run $X_1, \ldots, X_{50}$; see Fig 1 for an example.



Now we need to estimate s, the mixing time. Generate 500 independent runs of length 50, and compute the covariance between X_n and X_{n+s} for different values of s. See Fig 2 for a graph of covariance of X_{25} with X_n for $n = 1, \ldots, 50$.



We see that the covariance becomes small for COV(X_{25}, X_{30}), so we take $s = 5$. Now we generate a long run of length 1000, $X_1, \ldots, X_{1000}$. Then our estimate for the expected value is

$$\mathbb{E}[X^2] \simeq \frac{1}{200} \sum_{n=1}^{200} X_{5n}^2$$

A sample run with $K = 20$, random pdf for X, use MCMC to generate $X_1, \ldots, X_{1000}$ gives

$$\begin{aligned} \mathbb{E}[X^2] &\simeq 158.3 \quad \text{(MCMC approximation)} \\ \mathbb{E}[X^2] &= 154.8 \quad \text{(exact)} \end{aligned}$$

Example 2: sampling from a uniform distribution

Many interesting problems require computing the average of functions over complicated discrete sets. Such averages may be estimated by drawing independent samples from the set (IID samples) and computing the empirical average of the function over these samples. Namely, let Ω be a finite discrete set, and $f : \Omega \to \mathbb{R}$. Suppose we want to estimate the average

$$\mathbb{E}[f] = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} f(\omega)$$

Let $X_1, \ldots, X_k$ be drawn randomly and uniformly from Ω, and define

$$\overline{f}_k = \frac{1}{k} \sum_{i=1}^{k} f(X_i)$$

Then

$$\lim_{k \to \infty} \overline{f}_k = \mathbb{E}[f] \qquad \text{a.s.}$$

where the expectation on the right side is taken with respect to the uniform distribution on Ω, and the convergence is almost sure.

As a concrete example, consider the famous *knapsack problem*. Let $z = (z_1, \ldots, z_m)$ where each $z_i \in \{0, 1\}$, so $z \in \{0, 1\}^m$. Let $\{w_i\}$ be positive scalars, and $b > 0$. Define the discrete set

$$\Omega = \left\{ z \in \{0, 1\}^m \mid \sum_{j=1}^{m} w_j z_j \leq b \right\}$$

Assuming that $\sum_j w_j > b$, it follows that Ω is a proper (and complicated) subset of the hypercube $\{0, 1\}^m$. We want to sample randomly and uniformly from Ω. We do this by using the MCMC method to construct an appropriate Markov chain. Since the target distribution is uniform, the Markov chain should be symmetric. The update rule is computed as follows: given that $X_k = z = (z_1, \ldots, z_m)$ let $J \in \{1, \ldots, m\}$ be a uniform random variable (independent of all other r.v.'s) and define

$$Y = (z_1, \ldots, 1 - z_J, \ldots, z_m)$$

In other words, we flip the Jth bit to get the new point. Then we define

$$X_{k+1} = \begin{cases} Y & \text{if } Y \in \Omega \\ X_k & \text{if } Y \notin \Omega \end{cases}$$

The chain is clearly irreducible since it is possible to reach the state $(0, \ldots, 0)$ starting from any state. Also, since Ω is a proper subset of the hypercube, there is a state z and a choice of J such that $Y \notin \Omega$ and hence $X_{k+1} = X_k$; therefore the chain is aperiodic. Thus the chain is ergodic, and so converges to its unique stationary distribution. Finally the following computation shows that the chain is symmetric: let $w, z \in \Omega$ with $w \neq z$ such that there is $j \in \{1, \ldots, m\}$ with

$$\begin{aligned} w_i &= z_i, \quad i \neq j \\ w_j &= 1 - z_j \end{aligned} \tag{1}$$

Hence

$$\mathbb{P}(X_{k+1} = w | X_k = z) = \mathbb{P}(J = j) = \frac{1}{m}$$

Similarly

$$\mathbb{P}(X_{k+1} = z | X_k = w) = \mathbb{P}(J = j) = \frac{1}{m}$$

so the transition matrix is symmetric for this pair. Furthermore the transition matrix between states $w \neq z \in \Omega$ is nonzero if and only if (1) holds for some j, so this shows that the transition matrix is symmetric, and hence that the stationary distribution is uniform.

We can now use this Markov chain to estimate the average of a function on Ω. Consider for example the function

$$f(z) = e^{\beta \sum_j v_j z_j}$$

where $\beta > 0$ and $\{v_j\}$ is a fixed vector. The average is

$$\mathbb{E}[f] = \frac{1}{|\Omega|} \sum_{z \in \Omega} e^{\beta \sum_j v_j z_j}$$

(This is similar to the partition function encountered in statistical mechanics, where β is the inverse temperature.) By running the Markov chain, we generate samples X_k drawn uniformly from Ω. Let

$$Y_n = X_{N_0 + n N_1}, \quad n = 1, 2, \ldots$$

where N_0, N_1 are large integers. Our estimate of $\mathbb{E}[f]$ is

$$\widehat{f}_K = \frac{1}{K} \sum_{n=1}^{K} f(Y_n)$$

By choosing N_0 large we ensure that the marginal distribution of each Y_n is close to uniform, so

$$\mathbb{E}[\widehat{f}_K] \simeq \mathbb{E}[f]$$

By choosing N_1 large we ensure that the different Y_n are close to being independent, and hence

$$\mathrm{VAR}[\widehat{f}_K] \simeq \frac{1}{K} \mathrm{VAR}[f(Y)]$$

Example 3: Maximizing a function

Continuing with the knapsack problem, we may wish to find the maximum value of f over Ω. To do this, we construct a new MCMC algorithm whose stationary distribution is

$$\pi_\beta(z) = \frac{1}{\mathbb{E}[f]} f(z) = \frac{1}{\mathbb{E}[f]} e^{\beta \sum_j v_j z_j}$$

For large β this distribution is tightly concentrated around the value of z where f attains its maximum. Thus by sampling from π with large β we will with high probability generate a value of z which is close to the maximal value. We modify the MCMC construction from the previous section as follows: given that $X_k = z = (z_1, \ldots, z_m)$ let $J \in \{1, \ldots, m\}$ be a uniform random variable (independent of all other r.v.'s) and define

$$Y = (z_1, \ldots, 1 - z_J, \ldots, z_m)$$

If $Y \notin \Omega$ then $X_{k+1} = X_k$. If $Y \in \Omega$ then let

$$\alpha = \min \left\{ 1, \frac{\pi_\beta(Y)}{\pi_\beta(X_k)} \right\} = \min \left\{ 1, e^{\beta v_J (1 - 2 z_J)} \right\}$$

and define

$$X_{k+1} = \begin{cases} Y & \text{with probability } \alpha \\ X_k & \text{with probability } 1 - \alpha \end{cases}$$

Notice that the rejection probability increases as β increases, and this slows down the convergence of the algorithm. Simulated annealing is the method of slowly increasing β as the chain progresses, thereby focusing in on the region of high probability where we are most interested.

Example 4: Bayesian Inference

Let θ be a parameter in a family of probability distributions. Given a collection of data D, our goal is to find the best estimate of θ using this data. The Bayesian method requires the choice of an initial or prior distribution for θ, let us denote this by $\mathbb{P}_0$. Then the Bayesian formula for the posterior distribution $\mathbb{P}_1$ is

$$\mathbb{P}_1(\theta|D) = \frac{\mathbb{P}(D|\theta)\,\mathbb{P}_0(\theta)}{\int \mathbb{P}(D|\theta')\,d\mathbb{P}_0(\theta')}$$

where $\mathbb{P}(D|\theta)$ is the *likelihood* and $\int \mathbb{P}(D|\theta')\,d\mathbb{P}_0(\theta')$ is the *evidence*. Often the likelihood function and the prior can be easily evaluated but the evidence is hard to compute. This is a situation where MCMC can be very helpful.

Here is a concrete example. Suppose that at a party each entering guest is given a card with a number on it. The cards are numbered $1, 2, 3, \dots$ and they are handed out in sequence to entering guests. At the end of the evening many guests have left, but the host cannot remember how many cards were handed out, although she knows that the number was between 100 and 200. There are 20 people left in the room and she records the numbers on their cards. Given this information she tries to estimate the total number of guests. How should she do this?

The answer is to use Bayesian inference. Let M be the (unknown) number of guests, and let $x_1, \dots, x_k$ be the numbers on the cards of the remaining guests ($k = 20$ in this case). So as random variables the data is $D = (X_1, \dots, X_k)$ and the parameter is M. The prior for M is uniform on the interval $(n_1, n_2]$ (in this case $n_1 = 100$ and $n_2 = 200$). The likelihood is

$$\mathbb{P}(D|M = m) = \prod_{i=1}^{k} \mathbb{P}(X_i = x_i|M = m) = \prod_{i=1}^{k} \frac{1}{m}\theta(x_i \leq m)$$

where $\theta(A)$ is the indicator function for the event A:

$$\theta(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false} \end{cases}$$

Hence

$$\mathbb{P}(D|M = m) = \frac{1}{m^k}\,\theta(\max_i x_i \leq m)$$

The prior for M is

$$\mathbb{P}_0(M = m) = \frac{1}{n_2 - n_1}\,\theta(n_1 < m \leq n_2)$$

and so the evidence is

$$Z = \sum_m \mathbb{P}(D|M = m)\mathbb{P}_0(M = m) = \frac{1}{n_2 - n_1} \sum_{m=\max\{\max x_i, n_1+1\}}^{n_2} m^{-k}$$

Hence the posterior probability distribution for M is

$$\mathbb{P}_1(M = m|D) = \frac{1}{Z}\,\frac{1}{n_2 - n_1}\, m^{-k}\,\theta(\max\{\max x_i, n_1 + 1\} \le m \le n_2)$$

Note that one popular statistic is the *maximum likelihood estimator* $\widehat{M}_{MLE}$. This is the value of m which maximizes the posterior probability, that is the value which makes it *most likely that the given data would have been observed*. Clearly the MLE here is the maximum of the observed values, that is

$$\widehat{M}_{MLE} = \max x_i$$

However we might also be interested in the *mean* of the posterior, that is

$$\overline{M} = \sum_m m\,\mathbb{P}_1(M = m|D)$$

In this case we must evaluate the evidence Z, which of course is difficult to do analytically. So we use MCMC to simulate the distribution. Define

$$a = \max\{\max x_i, n_1 + 1\}, \quad b = n_2$$

The state space is $\Omega = \{a, a + 1, \ldots, b\}$ and we use a nearest neighbor model for our basic chain. Then the transition matrix is

$$\begin{aligned} q_{i,i+1} &= q_{i,i-1} = \frac{1}{2} \quad \text{for } a \le i \le b \\ q_{a,a} &= q_{b,b} = \frac{1}{2} \end{aligned}$$

This chain is aperiodic and irreducible. The MCMC transition matrix is then

$$\begin{aligned} p_{i,i+1} &= \frac{1}{2}\left(\frac{i+1}{i}\right)^{-k} \qquad i = a, a+1, \ldots, b-1 \\ p_{i,i-1} &= \frac{1}{2} \quad i = a+1, \ldots, b \end{aligned}$$

So the MCMC chain is a random walk on the integers with bias to the left.

As an example, suppose that $n_1 = 100$, $n_2 = 200$, $k = 20$ and $\max x_i = 140$. Then $a = 140$ and $b = 200$. We construct the Markov chain using the transition matrix above, so for example

$$p_{150,151} = 0.4378$$

We want to compute the mean of the posterior distribution, so we approximate this by

$$\overline{X} = \frac{1}{K}\sum_{i=1}^{K} Y_i$$

where $\{Y_i\}$ are samples from the chain. We want these to be approximately independent, and also approximately stationary. So we choose a random initial state X_0, and then generate $X_1, \ldots, X_N$ where N is chosen large enough so that the initial condition is forgotten. Then we set $Y_1 = X_N$. We continue, and generate $Y_2 = X_{2N}, Y_3 = X_{3N}, \ldots, Y_K = X_{KN}$. Using $N = K = 10,000$ we find for this example that

$$\overline{X} \simeq 147.2 \pm 0.2$$