11: Basics of Markov chain Monte Carlo Simulation

Kai Tan

University of Kentucky

Section 1

Introduction

Intro to Markov chain Monte Carlo (MCMC)

- **Goal**: Sample from p (eg, $p(\theta|y)$), or approx $E(f(X))$, $X \sim p$. Recall that $p(\theta|y)$ is very complicated and hard to sample from.
- **Procedure**: Start from θ_0, generate sequence $\theta_1, \theta_2, \ldots$. For each t, draw θ_t from a transition probability $T_t(\theta^t|\theta^{t-1})$ that depends only on the previous draw θ^{t-1}.
- **Key**: The transition probability distributions T_t must be constructed so that the Markov chain converges to a unique stationary distribution that is the posterior distribution $p(\theta|y)$.

Rationale for the Markov Chain simulation

- Markov chain simulation is to create a Markov process whose stationary distribution is the specified $p(\theta|y)$.
- Run the simulation long enough such that the distribution of the current draws is close enough to this stationary distribution.
- Good news: a variety of Markov chains with the desired property can be constructed.
  - Gibbs
  - Metropolis - Hastings
- Always check the convergence of the simulated sequences.

Section 2

Gibbs Sampler

The Gibbs Sampler

Idea: alternating conditional sampling. If $\theta = (\theta_1, \theta_2)$. The Gibbs sampler alternates between

1. $\theta_1^t \sim p(\theta_1 \mid \theta_2^{t-1}, y)$
2. $\theta_2^t \sim p(\theta_2 \mid \theta_1^t, y)$

If there are more parameters: $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$. The Gibbs sampler alternates between

1. $\theta_1^t \sim p(\theta_1 \mid \theta_{2:d}^{t-1}, y)$
2. $\theta_2^t \sim p(\theta_2 \mid \theta_1^t, \theta_{3:d}^{t-1}, y)$
3. $\vdots$
4. $\theta_d^t \sim p(\theta_d \mid \theta_{1:d-1}^t, y)$

This is only possible if you can sample from the **conditional posteriors** (i.e. need conditional conjugacy).

The Gibbs Sampler: a toy example I

- Bivariate normal distribution

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \tag{1}$$

- Uniform prior $p(\theta_1, \theta_2) = U(0,1) \times U(0,1)$.
- A single obs. (y_1, y_2)
- Target (posterior) distribution:

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \Big| y \sim N_2 \left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{2}$$

We can derive the conditional posterior distributions

1. $\theta_1 \mid \theta_2, y \sim \text{Normal}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$
2. $\theta_2 \mid \theta_1, y \sim \text{Normal}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$

The Gibbs Sampler: a toy example II

$\rho = 0.8$, $(y_1, y_2) = (0, 0)$, initial values: $\theta^0 = (\pm 2.5, \pm 2.5)$

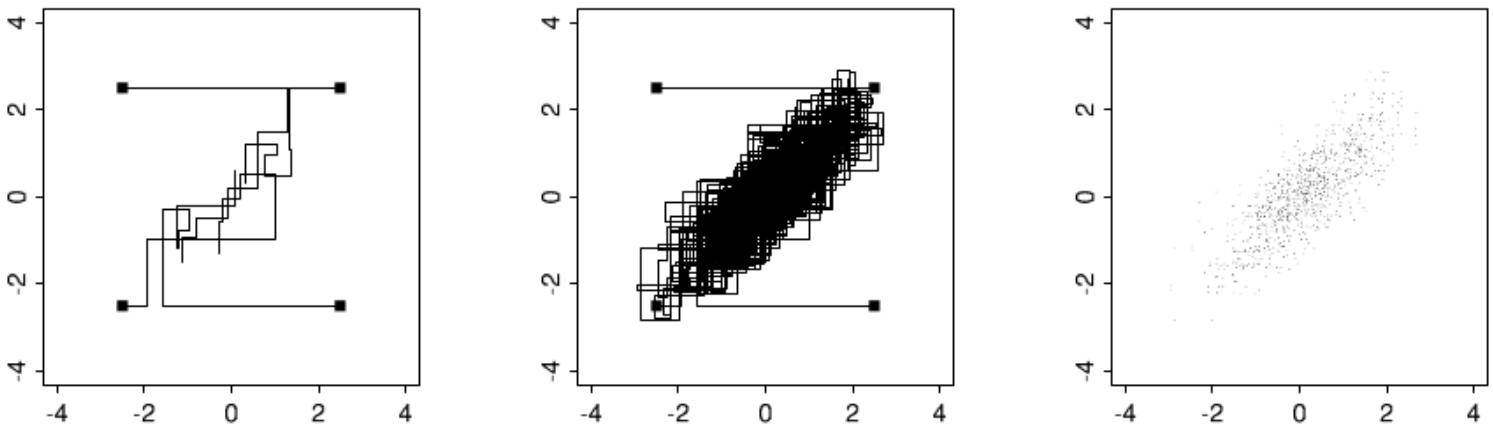

Figure 11.2 *Four independent sequences of the Gibbs sampler for a bivariate normal distribution with correlation $\rho = 0.8$, with overdispersed starting points indicated by solid squares. (a) First 10 iterations, showing the componentwise updating of the Gibbs iterations. (b) After 500 iterations, the sequences have reached approximate convergence. Figure (c) shows the points from the second halves of the sequences, representing a set of correlated draws from the target distribution.*

Section 3

Metropolis and Metropolis-Hastings algorithm

Metropolis I

- Draw a starting point θ^0 from $p(\theta^0)$, s.t. $p(\theta^0|y) > 0$
- For $t = 1, 2, \ldots,$
  1. Sample a proposal θ^* from a jumping distribution $J_t(\theta^*|\theta^{t-1})$ at time t.
     J is **symmetric**: $J(\theta_a|\theta_b) = J(\theta_b|\theta_a)$ for all θ_a, θ_b at any t.
  2. Calculate the ratio of the targeted posterior densities,

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}.$$

  3. Set

$$\theta^t = \begin{cases} \theta^* & \textit{with prob. } min(r, 1) \\ \theta^{t-1} & \textit{otherwise.} \end{cases}$$

That is, we accept our proposal θ^* with prob. $\min(r, 1)$.
Make sense: the larger r, the larger acceptance prob.

Metropolis II

- It requires to calculate the ratio $r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}$.

Note that $p(\theta|y) = \frac{q(\theta|y)}{\int q(\theta|y)d\theta}$, where $q(\theta|y)$ is the unnormalized posterior density. Thus,

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)} = \frac{q(\theta^*|y)}{q(\theta^{t-1}|y)}.$$

Don't need to know the normalizing constant.

Relation to optimization

Acceptance/rejection rule:

- If the jump increases the posterior density, i.e. $r > 1$, set $\theta^t = \theta^*$.
- If the jump decreases the posterior density, i.e. $r < 1$, set $\theta^t = \theta^*$ with prob. r. Otherwise $\theta^t = \theta^{t-1}$.
- How to implement above steps?
  - Generate a random number $u \sim U(0, 1)$
  - Set $\theta^t = \theta^*$, if $u \leq r$. (The prob. of this event is r)
  - Set $\theta^t = \theta^{t-1}$, if $u > r$.

Metropolis: Example I

- Target: $p(\theta|y) = N_2(\theta|0, I)$
- Jumping dist.: $J_t(\theta^*|\theta^{t-1}) = N_2(\theta^*|\theta^{t-1}, 0.2^2 I)$, i.e., sample θ^* from $N_2(\theta^{t-1}, 0.2^2 I)$. Easy to check that J is symmetric.
- Density ratio (acceptance rate)

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)} = \frac{f(\theta^*)}{f(\theta^{t-1})},$$

where f is the pdf for $N_2(0, I)$.

Metropolis: Example II

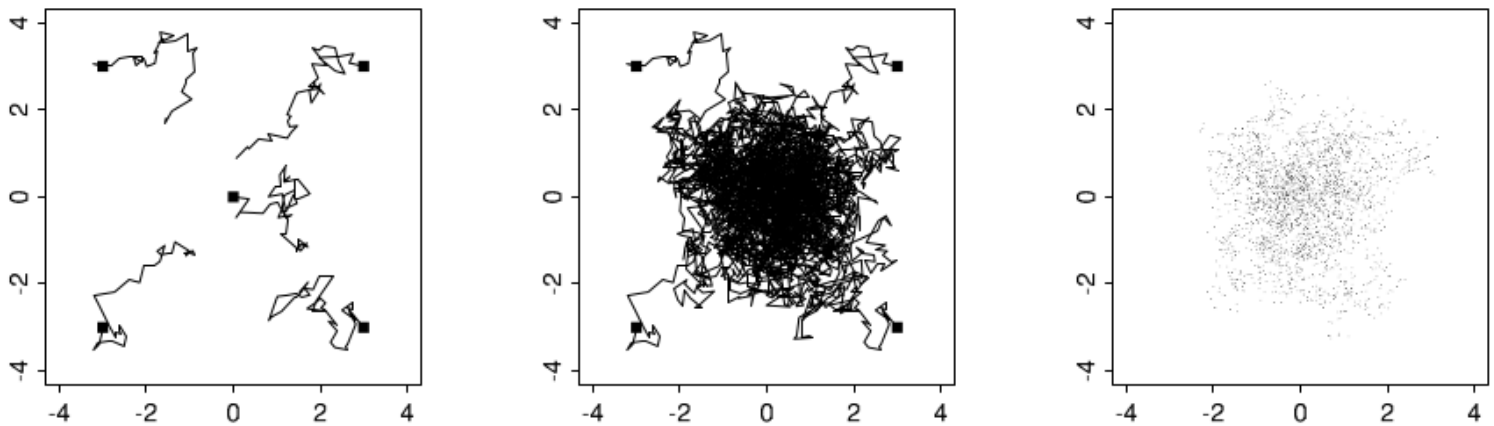

Figure 11.1 *Five independent sequences of a Markov chain simulation for the bivariate unit normal distribution, with overdispersed starting points indicated by solid squares. (a) After 50 iterations, the sequences are still far from convergence. (b) After 1000 iterations, the sequences are nearer to convergence. Figure (c) shows the iterates from the second halves of the sequences; these represent a set of (correlated) draws from the target distribution. The points in Figure (c) have been jittered so that steps in which the random walks stood still are not hidden. The simulation is a Metropolis algorithm described in the example on page 278, with a jumping rule that has purposely been chosen to be inefficient so that the chains will move slowly and their random-walk-like aspect will be apparent.*

Why Metropolis works? I

To show the sequence $\theta^1, \theta^2, \ldots,$ converges to the target distribution.

i) The Markov chain has a unique stationary distribution. (irreducible + aperiodic + not transient.)
   - irreducible: it is possible to get to any state from any state.
   - aperiodic: $k = \gcd\{n > 0 : \Pr(X_n = i \mid X_0 = i) > 0\}$. $k = 1$.
   - not transient: $\sum_{n=1}^{\infty} \mathbb{P}(X_n = i \mid X_0 = i) = \infty$.

ii) The stationary distribution equals the target distribution.
   - Consider at time $t-1$ with a draw θ^{t-1} from the target dist. $p(\theta|y)$. Two points θ_a and θ_b, s.t. $p(\theta_b|y) > p(\theta_a|y)$.
   - From θ_a to θ_b:

$$\begin{aligned} p(\theta^{t-1} = \theta_a, \theta^t = \theta_b) &= p(\theta_a|y) J_t(\theta_b|\theta_a) \min(r, 1) \\ &= p(\theta_a|y) J_t(\theta_b|\theta_a). \end{aligned}$$

The last equality is due to the fact $r = 1$, which is by $p(\theta_b|y) > p(\theta_a|y)$.

Why Metropolis works? II

- From θ_b to θ_a:

$$\begin{aligned} p(\theta^{t-1} = \theta_b, \theta^t = \theta_a) &= p(\theta_b|y) J_t(\theta_a|\theta_b) \min(r, 1) \\ &= p(\theta_b|y) J_t(\theta_a|\theta_b) \frac{p(\theta_a|y)}{p(\theta_b|y)} \\ &= p(\theta_a|y) J_t(\theta_a|\theta_b) \\ &= p(\theta_a|y) J_t(\theta_b|\theta_a) \quad \text{(J is symmetric)} \\ &= p(\theta^{t-1} = \theta_a, \theta^t = \theta_b). \end{aligned}$$

- The joint distribution $f_{\theta^{t-1},\theta^t}(x_1, x_2)$ is symmetric, θ^{t-1}, θ^t has the same marginal distribution, and so $p(\theta|y)$ is the stationary distribution.
- $\sum_{\theta_a} p(\theta_a|y) J_t(\theta_b|\theta_a) = \sum_{\theta_a} p(\theta^{t-1} = \theta_a, \theta^t = \theta_b) = p(\theta_b|y)$. ($\pi$ is stationary dist. : $\pi P = \pi$, *or* $\sum_i \pi_i P_{ij} = \pi_j$).

Metropolis-Hastings

- A generalization of Metropolis algorithms.
- Do not require jumping dist. J_t to be symmetric.
- Adjust our acceptance rate r due to the asymmetric as a ratio of ratios:

$$r = \frac{p(\theta^*|y)/J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/J_t(\theta^{t-1}|\theta^*)}.$$

  Well-defined: as jumping happens only when $J_t \neq 0$.
- The asymmetric jumping dist. can be useful in increasing the speed of the random walk.
- Convergence to the target distribution is proved in the same way as for the Metropolis algorithm.

Metropolis-Hastings Algorithm

- Draw a starting point θ^0 from $p(\theta^0)$, s.t. $p(\theta^0|y) > 0$
- For $t = 1, 2, \ldots,$
  1. Sample a proposal θ^* from a jumping distribution $J_t(\theta^*|\theta^{t-1})$at time t. J_t can be asymmetric.
  2. Calculate the ratio of the posterior densities,

$$r = \frac{p(\theta^*|y)/J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/J_t(\theta^{t-1}|\theta^*)}.$$

  3. Set

$$\theta^t = \begin{cases} \theta^* & \textit{with prob. } min(r, 1) \\ \theta^{t-1} & \textit{otherwise.} \end{cases}$$

That is, we accept our proposal θ^* with prob. $\min(r, 1)$.
Make sense: the larger r, the larger acceptance prob.

Convergence of Metropolis-Hastings

- Consider at time $t-1$ with a draw θ^{t-1} from the target dist. $p(\theta|y)$.
  Two points θ_a and θ_b, $p(\theta_b|y)J_t(\theta_a|\theta_b) \geq p(\theta_a|y)J_t(\theta_b|\theta_a)$.
- From θ_a to θ_b:

$$\begin{aligned} p(\theta^{t-1} = \theta_a, \theta^t = \theta_b) &= p(\theta_a|y)J_t(\theta_b|\theta_a)\min(r,1) \\ &= p(\theta_a|y)J_t(\theta_b|\theta_a). \end{aligned}$$

  The last equality is due to the fact $r = 1$, which is by $p(\theta_b|y)J_t(\theta_a|\theta_b) \geq p(\theta_a|y)J_t(\theta_b|\theta_a)$.
- From θ_b to θ_a:

$$\begin{aligned} p(\theta^{t-1} = \theta_b, \theta^t = \theta_a) &= p(\theta_b|y)J_t(\theta_a|\theta_b)\min(r,1) \\ &= p(\theta_b|y)J_t(\theta_a|\theta_b)\frac{p(\theta_a|y)J_t(\theta_b|\theta_a)}{p(\theta_b|y)J_t(\theta_a|\theta_b)} \\ &= p(\theta_a|y)J_t(\theta_b|\theta_a) = p(\theta^{t-1} = \theta_a, \theta^t = \theta_b). \end{aligned}$$

Ideal jumping for Metropolis-Hastings

- Sample the proposal, θ^*, from the target distribution, $J(\theta^*|\theta) = p(\theta^*|y)$ for all θ.
- The ratio

$$\begin{aligned} r &= \frac{p(\theta^*|y)/J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/J_t(\theta^{t-1}|\theta^*)} \\ &= \frac{p(\theta^*|y)/p(\theta^*|y)}{p(\theta^{t-1}|y)/p(\theta^{t-1}|y)} \\ &= 1. \end{aligned}$$

- The iterates θ^t are a sequence of independent draws from $p(\theta|y)$.

Not practical, because (we assume) it's hard to sample from $p(\theta|y)$.

How to choose a good jumping distribution

Four criteria:

- For any θ, it is easy to sample from $J(\theta^*|\theta)$.
- It is easy to compute the ratio r.
- Each jump goes a reasonable distance in the parameter space (otherwise the random walk moves too slowly).
- The jumps are not rejected too frequently (otherwise the random walk wastes too much time standing still).

The topic of constructing efficient simulation algorithms is in the chapter 12.

Another look at the Gibbs sampler

Gibbs sampling is a special case of the Metropolis-Hastings algorithm.

- single updated (or blocked)
- proposal distribution is the conditional distribution
  - proposal and target distributions are the same
  - acceptance probability is 1
- $\theta = (\theta_1, \ldots, \theta_d)$. For $j = 1, \ldots, d$,

$$J_{j,t}^{\text{Gibbs}}(\theta^*|\theta^{t-1}) = \begin{cases} p(\theta^*|\theta_{-j}^{t-1}, y) & \text{if } \theta_{-j}^* = \theta_{-j}^{t-1}; \\ 0 & \text{otherwise.} \end{cases}$$

  The only possible jumps are to parameter vectors θ^* that match θ^{t-1} on all components other than the jth.

- $$r = \frac{p(\theta^*|y)/J_{j,t}^{\text{Gibbs}}(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/J_{j,t}^{\text{Gibbs}}(\theta^{t-1}|\theta^*)} = \frac{p(\theta^*|y)/p(\theta_j^*|\theta_{-j}^{t-1},y)}{p(\theta^{t-1}|y)/p(\theta_j^{t-1}|\theta_{-j}^{t-1},y)} =$$
$$\frac{p(\theta_j^*,\theta_{-j}^{t-1}|y)/p(\theta_j^*|\theta_{-j}^{t-1},y)}{p(\theta_j^{t-1},\theta_{-j}^{t-1}|y)/p(\theta_j^{t-1}|\theta_{-j}^{t-1},y)} = \frac{p(\theta_{-j}^{t-1}|y)}{p(\theta_{-j}^{t-1}|y)} = 1.$$

summary

- Markov chain Monte Carlo techniques produce correlated draws from the posterior of interest.
- Three sampler: Gibbs, Metropolis, and Metropolis-Hastings.
- The Markov chain converges to a unique stationary distribution that is the targeted posterior distribution.

Next Class

- Inference and assessing convergence
- Effective number of simulation draws
- More examples on hierarchical normal model.