Simulation - Lectures - Part III
Markov chain Monte Carlo

Julien Berestycki

Part A Simulation and Statistical Programming

Hilary Term 2018

Outline

Markov Chain Monte Carlo

Recap on Markov chains

Markov Chain Monte Carlo

Metropolis-Hastings

Markov chain Monte Carlo Methods

- Our aim is to estimate $\theta = \mathbb{E}_p(\phi(X))$ where X is a random variable on Ω with pmf (or pdf) p.
- Up to this point we have based our estimates on independent and identically distributed draws from either p itself, or some proposal distribution with pmf/pdf q.
- In MCMC we simulate a correlated sequence $X_0, X_1, X_2, ...$ such that X_t is approximately distributed from p for t large, and rely on the usual estimate

$$\hat{\theta}_n = \frac{1}{n}\sum_{t=0}^{n-1}\phi(X_t).$$

- We will suppose the space of states of X is discrete (finite or countable).
- But it should be kept in mind that MCMC methods are applicable to continuous state spaces, and in fact one of the most versatile and widespread classes of Monte Carlo algorithms currently.

Outline

Markov Chain Monte Carlo

- Recap on Markov chains
- Markov Chain Monte Carlo
- Metropolis-Hastings

Markov chains

- From Part A Probability.
- Let $(X_t)_{t=0,1,\dots}$ be a homogeneous Markov chain of random variables on Ω with starting distribution $X_0 \sim p^{(0)}$ and transition matrix $P = (P_{ij})_{i,j\in\Omega}$ with

$$P_{i,j} = \mathbb{P}(X_{t+1} = j | X_t = i).$$

  for $i, j \in \Omega$.
- We write $(X_t)_{t=0,1,\dots} \sim \mathrm{Markov}(p^{(0)}, P)$
- Denote by $P^{(n)}_{i,j}$ the n-step transition probabilities

$$P^{(n)}_{i,j} = \mathbb{P}(X_{t+n} = j | X_t = i)$$

  and by $p^{(n)}(i) = \mathbb{P}(X_n = i)$.
- Recall that a Markov chain, or equivalently its transition matrix P, is irreducible if and only if, for each pair of states $i, j \in \Omega$ there is n such that $P^{(n)}_{i,j} > 0$.

Markov chains

- π is a stationary, or invariant distribution of P, if π verifies

  $$\pi_j = \sum_{i \in \Omega} \pi_i P_{ij}$$

  for all $j \in \Omega$.
- If $p^{(0)} = \pi$ then

  $$p^{(1)}(j) = \sum_{i \in \Omega} p^{(0)}(i) P_{i,j},$$

  so $p^{(1)}(j) = \pi(j)$ also. Iterating, $p^{(t)} = \pi$ for each $t = 1, 2, ...$ in the chain, so the distribution of X_t doesn't change with t, it is stationary.
- If P is irreducible, and has a stationary distribution π, then this stationary distribution π is unique.

Ergodic Theorem for Markov chains

Theorem (Theorems 6.1, 6.2, 6.3 of Part A Probability)

Let $(X_t)_{t=0,1,\ldots} \sim \mathrm{Markov}(\lambda, P)$ be an irreducible Markov chain on a discrete state space Ω. Assume it admits a stationary distribution π. Then, for any initial distribution λ

$$\frac{1}{n}\sum_{t=0}^{n-1} \mathbb{I}(X_t = i) \to \pi(i) \quad \textit{almost surely, as } n \to \infty.$$

That is

$$\mathbb{P}\left(\frac{1}{n}\sum_{t=0}^{n-1} \mathbb{I}(X_t = i) \to \pi(i)\right) = 1.$$

Additionally, if the chain is aperiodic, then for all $i \in \Omega$

$$\mathbb{P}(X_n = i) \to \pi(i)$$

Ergodic Theorem for Markov chains

Corollary

Let $(X_t)_{t=0,1,\ldots} \sim \mathrm{Markov}(\lambda, P)$ be an irreducible Markov chain on a discrete state space Ω. Assume it admits a stationary distribution π. Let $\phi : \Omega \to \mathbb{R}$ be a bounded function, X a discrete random variable on Ω with pmf π and $\theta = \mathbb{E}_p[\phi(X)] = \sum_{i \in \Omega} \phi(i)\pi(i)$. Then, for any initial distribution λ

$$\frac{1}{n}\sum_{t=0}^{n-1} \phi(X_t) \to \theta \quad \text{almost surely, as } n \to \infty.$$

That is

$$\mathbb{P}\left(\frac{1}{n}\sum_{t=0}^{n-1} \phi(X_t) \to \theta\right) = 1.$$

Ergodic Theorem for Markov chains

Proof (non-examinable). Assume wlog $|\phi(i)| < 1$ for all $i \in \Omega$. Let $A \subset \Omega$ and $V_i(n) = \sum_{t=0}^{n-1} \mathbb{I}(X_t = i)$.

$$\begin{aligned}
\left|\frac{1}{n}\sum_{t=0}^{n-1}\phi(X_t) - \theta\right| &= \left|\frac{1}{n}\sum_{t=0}^{n-1}\sum_{i\in\Omega}\phi(i)\mathbb{I}(X_t = i) - \sum_{i\in\Omega}\phi(i)\pi(i)\right| \\
&= \left|\sum_{i\in\Omega}\phi(i)\left(\frac{1}{n}V_i(n) - \pi(i)\right)\right| \\
&\le \sum_{i\in A}\left|\frac{V_i(n)}{n} - \pi(i)\right| + \sum_{i\notin A}\left|\frac{V_i(n)}{n} - \pi(i)\right| \\
&\le \sum_{i\in A}\left|\frac{V_i(n)}{n} - \pi(i)\right| + \sum_{i\notin A}\left(\frac{V_i(n)}{n} + \pi(i)\right) \\
&= \sum_{i\in A}\left|\frac{V_i(n)}{n} - \pi(i)\right| + \sum_{i\in A}\left(\pi(i) - \frac{V_i(n)}{n}\right) + 2\sum_{i\notin A}\pi(i) \\
&\le 2\sum_{i\in A}\left|\frac{V_i(n)}{n} - \pi(i)\right| + 2\sum_{i\notin A}\pi(i)
\end{aligned}$$

where in line 5 we've used $\sum_{i\notin A}\frac{V_i(n)}{n} = 1 - \sum_{i\in A}\frac{V_i(n)}{n} = \sum_i \pi(i) - \sum_{i\in A}\frac{V_i(n)}{n}$.

Ergodic Theorem for Markov chains

Proof (continued). Let $\epsilon > 0$, and take A finite such that $\sum_{i \notin A} \pi(i) < \epsilon/4$. For $N \in \mathbb{N}$, Define the event

$$E_N = \left\{ \sum_{i \in A} \left| \left(\frac{V_i(n)}{n} - \pi(i) \right) \right| < \epsilon/4 \text{ for all } n \geq N \right\}.$$

As $\mathbb{P}(\frac{V_i(n)}{n} \to \pi(i)) = 1$ for all $i \in \Omega$ and A is finite, the event E_N must occur for some N hence $\mathbb{P}(\cup E_N) = 1$. It follows that, for any $\epsilon > 0$

$$\mathbb{P}\left(\exists N \text{ such that for all } n \geq N, \quad \left| \frac{1}{n} \sum_{t=0}^{n-1} \phi(X_t) - \theta \right| < \epsilon \right) = 1.$$

Reversible Markov chains

- In a reversible Markov chain we cannot distinguish the direction of simulation from inspection of a realization of the chain and its reversal, even with knowledge of the transition matrix.
- A Markov chain $(X_t)_{t\geq 0} \sim \mathrm{Markov}(\pi, P)$ is reversible iff

  $$\mathbb{P}(X_0, X_1, \ldots, X_n) = \mathbb{P}(X_n, X_{n-1}, \ldots, X_0)$$

  for any $n \geq 0$.
- We also say that P is reversible with respect to π, or π-reversible
- Most MCMC algorithms are based on reversible Markov chains.

Reversible Markov chains

- It seems clear that a Markov chain will be reversible if and only if $\mathbb{P}(X_{t-1} = j | X_t = i) = P_{i,j}$, so that any particular transition occurs with equal probability in forward and reverse directions.

Theorem

Let P be a transition matrix. If there is a probability mass function π on Ω such that π and P satisfy the detailed balance condition

$$\pi(i)P_{i,j} = \pi(j)P_{j,i} \qquad \text{for all pairs } i, j \in \Omega,$$

then

(I) $\pi = \pi P$, so π is stationary for P and
(II) the chain $(X_t) \sim \mathrm{Markov}(\pi, P)$ is reversible.

Reversible Markov chains

- Proof of (I): sum both sides of detailed balance equation over $i \in \Omega$. Now $\sum_i P_{j,i} = 1$ so $\sum_i \pi(i) P_{i,j} = \pi(j)$.
- Proof of (II), we have π a stationary distribution of P so $\mathbb{P}(X_t = i) = \pi(i)$ for all $t = 1, 2, \ldots$ along the chain. Then

$$\begin{aligned}\mathbb{P}(X_{t-1} = j | X_t = i) &= \mathbb{P}(X_t = i | X_{t-1} = j) \frac{\mathbb{P}(X_{t-1} = j)}{\mathbb{P}(X_t = i)} \text{ (Bayes rule)} \\ &= P_{j,i} \pi(j) / \pi(i) \text{ (stationarity)} \\ &= P_{i,j} \text{ (detailed balance).}\end{aligned}$$

Outline

Markov Chain Monte Carlo

- Recap on Markov chains
- Markov Chain Monte Carlo
- Metropolis-Hastings

Markov chain Monte Carlo method (discrete case)

Let X be a discrete random variable with pmf p on Ω, ϕ a bounded function on Ω and $\theta = \mathbb{E}_p[\phi(X)]$. Consider a homogeneous Markov chain $(X_t)_{t=0,1,\ldots} \sim \mathrm{Markov}(\lambda, P)$ with initial distribution λ on Ω and transition matrix P, such that P is irreducible, and admits p as invariant distribution. Then, for any initial distribution λ the MCMC estimator

$$\widehat{\theta}_n^{\mathtt{MCMC}} = \frac{1}{n} \sum_{i=1}^{n-1} \phi(X_t)$$

is (weakly and strongly) consistent

$$\widehat{\theta}_n^{\mathtt{MCMC}} \to \theta \text{ almost surely as } n \to \infty$$

and, if the chain is aperiodic

$$X_t \to p \text{ in distribution as } t \to \infty.$$

Markov chain Monte Carlo method

- Proof follows directly from the ergodic theorem and corollary
- Note that the estimator is biased, as $\lambda \neq p$ (otherwise we would use a standard Monte Carlo estimator)
- For t large, we have $X_t \overset{d}{\simeq} X$
- In order to implement the MCMC algorithm we need, for a given target distribution p, to find an irreducible (and aperiodic) transition matrix P with admits p as invariant distribution
- Most MCMC algorithms use a transition matrix P which is reversible with respect to p
- The Metropolis-Hastings algorithm provides a generic way to obtain such P for any target distribution p

Outline

Markov Chain Monte Carlo

Recap on Markov chains

Markov Chain Monte Carlo

Metropolis-Hastings

Metropolis-Hastings Algorithm

- The Metropolis-Hastings (MH) algorithm allows to simulate a Markov Chain with any given stationary distribution.
- We will start with simulation of random variable X on a discrete state space.
- Let $p(x) = \tilde{p}(x)/Z_p$ be the pmf on Ω. We will call p the (pmf of the) target distribution.
- To simplify notations, we assume that $p(x) > 0$ for all $x \in \Omega$
- Choose a 'proposal' transition matrix $q(y|x)$. We will use the notation $Y \sim q(\cdot|x)$ to mean $\Pr(Y = y|X = x) = q(y|x)$.

Metropolis-Hastings Algorithm

Metropolis-Hastings algorithm

```
1. Set the initial state $X_0 = x_0$.
2. For $t = 1, 2, \dots, n-1$:
    2.1 Assume $X_{t-1} = x_{t-1}$.
    2.2 Simulate $Y_t \sim q(\cdot|x_{t-1})$ and $U_t \sim \mathrm{U}[0,1]$.
    2.3 If
            $$U_t \leq \alpha(Y_t|x_{t-1})$$
        where
            $$\alpha(y|x) = \min\left\{1, \frac{\tilde{p}(y)q(x|y)}{\tilde{p}(x)q(y|x)}\right\}$$
        set $X_t = Y_t$, otherwise set $X_t = x_{t-1}$.
```

Metropolis-Hastings Algorithm

- The Metropolis-Hastings algorithm defines a Markov chain with transition matrix P such that, for $x, y \in \Omega$

$$\begin{aligned} P_{x,y} &= \mathbb{P}(X_t = y | X_{t-1} = x) \\ &= q(y|x)\alpha(y|x) + \rho(x)\mathbb{I}(y = x) \end{aligned}$$

where $\rho(x)$ is the probability of rejection

$$\rho(x) = 1 - \sum_{y \in \Omega} q(y|x)\alpha(y|x).$$

Metropolis-Hastings Algorithm

Theorem

The transition matrix P of the Markov chain generated by the Metropolis-Hastings algorithm is reversible with respect to p and therefore admits p as stationary distribution.

- Proof: We check detailed balance. For $x \neq y$

$$
\begin{aligned}
p(x)P_{x,y} &= p(x)q(y|x)\alpha(y|x) \\
&= p(x)q(y|x)\min\left\{1, \frac{p(y)q(x|y)}{p(x)q(y|x)}\right\} \\
&= \min\left\{p(x)q(y|x), p(y)q(x|y)\right\} \\
&= p(y)q(x|y)\min\left\{\frac{p(x)q(y|x)}{p(y)q(x|y)}, 1\right\} \\
&= p(y)q(x|y)\alpha(x|y) \\
&= p(y)P_{y,x}.
\end{aligned}
$$

Metropolis-Hastings Algorithm

- To run the MH algorithm, we need to specify $X_0 = x_0$ (or $X_0 \sim \lambda$) and a proposal $q(y|x)$.
- We only need to know the target p up to a normalizing constant as α depends only on $p(y)/p(x) = \tilde{p}(y)/\tilde{p}(x)$.
- If the Markov chain simulated by the MH algorithm is irreducible and aperiodic then the ergodic theorem applies.
- Verifying aperiodicity is usually straightforward, since the MCMC algorithm may reject the candidate state y, so $P_{x,x} > 0$ for at least some states $x \in \Omega$.
- In order to check irreducibility we need to check that q can take us anywhere in Ω (so q itself is an irreducible transition matrix), and then that the acceptance step doesn't trap the chain (as might happen if $\alpha(y|x)$ is zero too often).

Example: Discrete Distribution on a finite state-space

- Consider a discrete random variable $X \sim p$ on $\Omega = \{1, 2, ..., m\}$ with $\tilde{p}(i) = i$ so $Z_p = \sum_{i=1}^{m} i = \frac{m(m+1)}{2}$.
- One simple proposal distribution is $Y \sim q$ on Ω such that $q(i) = 1/m$.
- Acceptance probability

$$\alpha(y|x) = \min\left\{1, \frac{\tilde{p}(y)q(x|y)}{\tilde{p}(x)q(y|x)}\right\} = \min\left\{1, \frac{y}{x}\right\}$$

- This proposal scheme is clearly irreducible

$$\begin{aligned}\mathbb{P}(X_{t+1} = y|X_t = x) &\geq q(y|x)\alpha(y|x) \\ &= \frac{1}{m}\min(1, y/x) > 0\end{aligned}$$

Example: Discrete Distribution on a finite state-space

- Start from $X_0 = 1$.
- For $t = 1, \ldots, n-1$
  1. Let $Y_t \sim \mathrm{U}\{1, 2, ..., m\}$ and $U_t \sim \mathrm{U}[0, 1]$
  2. If
     $$U_t \leq \frac{Y_t}{X_{t-1}}$$
     set $X_t = Y_t$, otherwise set $X_t = X_{t-1}$.
- For t large, $X_t \overset{d}{\simeq} X$

Example: Discrete Distribution on a finite state-space

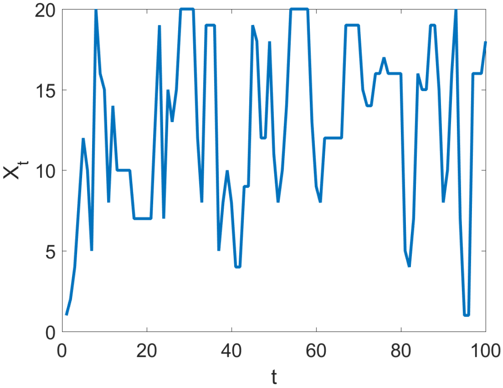


Figure: Realization of the MH Markov chain for $n = 100$ with $m = 20$.

Example: Discrete Distribution on a finite state-space

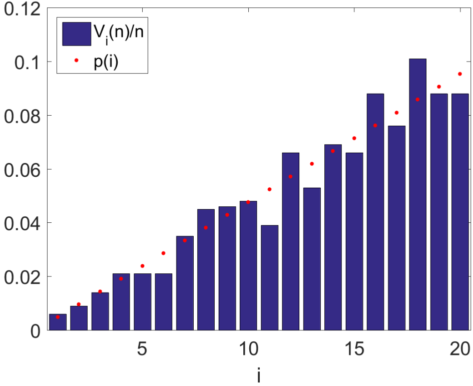


Figure: Average number of visits $V_i(n)/n$ ($n = 1000$) and target pmf $p(i)$

Example: Discrete Distribution on a finite state-space

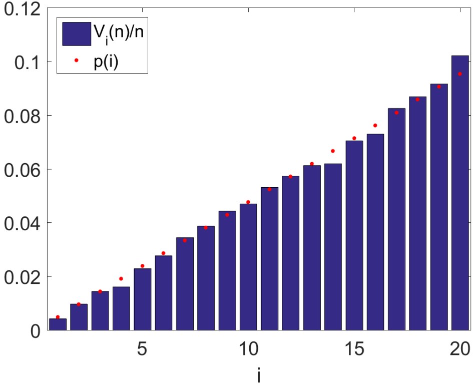


Figure: Average number of visits $V_i(n)/n$ ($n = 10,000$) and target pmf $p(i)$

Example: Poisson Distribution

- We want to simulate $p(x) = e^{-\lambda}\lambda^x/x! \propto \lambda^x/x!$
- For the proposal we use

$$q(y|x) = \begin{cases} \frac{1}{2} & \text{for } y = x \pm 1, x \geq 1 \\ 1 & \text{for } x = 0, y = 1 \\ 0 & \text{otherwise,} \end{cases}$$

  i.e. toss a coin and add or substract 1 to x to obtain y.
- Acceptance probability

$$\alpha(y|x) = \begin{cases} \min\left(1, \frac{\lambda}{x+1}\right) & \text{if } y = x + 1, x \geq 1 \\ \min\left(1, \frac{x}{\lambda}\right) & \text{if } y = x - 1, x \geq 2 \end{cases}$$

  and $\alpha(1|0) = \min(1, \lambda/2)$, $\alpha(0|1) = \min(1, 2/\lambda)$.
- Markov chain is irreducible (check!)

Example: Poisson Distribution

```
▶ Set $X_0 = 1$.
▶ For $t = 1, \dots, n-1$
    1. If $X_{t-1} = 0$, set $Y_t = 1$
    2. Otherwise, simulate $V_t \sim \mathrm{U}[0,1]$
        2.1 If $V_t \leq \frac{1}{2}$, set $Y_t = X_{t-1} + 1$.
        2.2 Otherwise set $Y_t = X_{t-1} - 1$.
    3. Simulate $U_t \sim \mathrm{U}[0,1]$.
    4. If $U_t \leq \alpha(Y_t|X_{t-1})$, set $X_t = Y_t$, otherwise set $X_t = X_{t-1}$.
```

Example: Poisson distribution

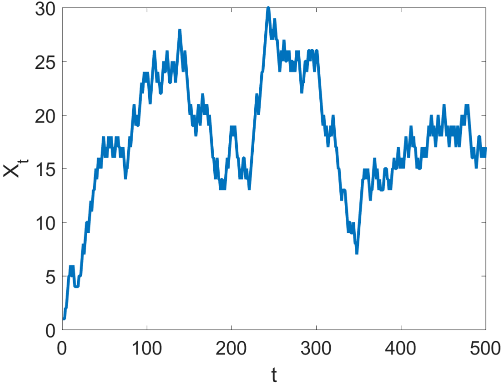


Figure: Realization of the MH Markov chain for $n = 500$ with $\lambda = 20$.

Example: Poisson distribution

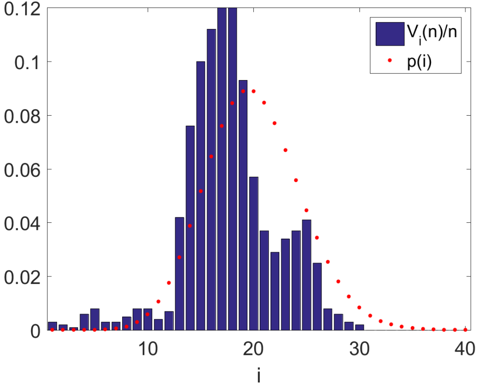


Figure: Average number of visits $V_i(n)/n$ ($n = 1000$) and target pmf $p(i)$

Example: Poisson distribution

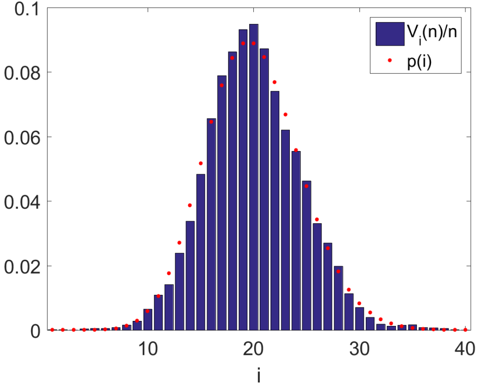


Figure: Average number of visits $V_i(n)/n$ ($n = 10,000$) and target pmf $p(i)$

Example: Image

- Consider a $m_1 \times m_2$ image, where $I(i,j) \in \{0, 1, \ldots, 256\}$ is the gray level of pixel $(i,j) \in \Omega = \{0, \ldots, m_1 - 1\} \times \{0, \ldots, m_2 - 1\}$
- Consider a discrete random variable taking values in Ω
- Unnormalized pdf

$$\tilde{p}((i,j)) = I(i,j)$$

- Proposal transition probabilities

$$q((y_1, y_2)|(x_1, x_2)) = q(y_1|x_1)q(y_2|x_2)$$

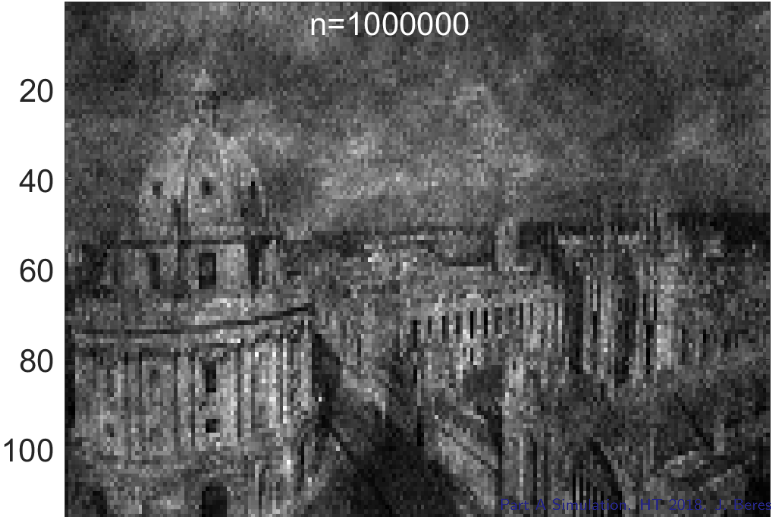
with

$$q(y_1|x_1) = \begin{cases} 1/3 & \text{if } y_1 = x_1 \pm 1 \text{ or } y_1 = x_1, \quad \mod m_1 \\ 0 & \text{otherwise} \end{cases}$$

and similarly for $q(y_2|x_2)$.

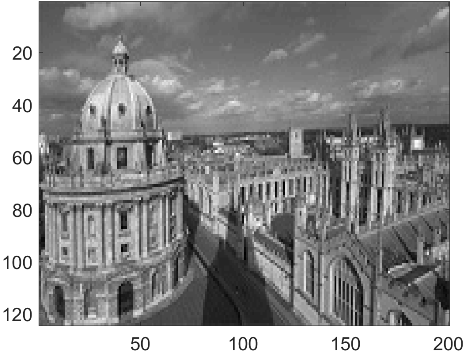
Example: Simulation of an image

- Average number of visits to each pixel (i, j):
  $V_{(i,j)}(n) = \frac{1}{n}\sum_{t=0}^{n-1} \mathbb{I}(X_t = (i, j))$



Example: Simulation of an image

- Target pmf



Metropolis-Hastings algorithm on $\mathbb{R}^d$

- The Metropolis-Hastings algorithm generalizes to continuous state-space where $\Omega \subseteq \mathbb{R}^d$ with
  1. p is a pdf on Ω
  2. $q(\cdot|x)$ is a pdf on Ω for any $x \in \Omega$
- The Metropolis-Hastings algorithm thus defines a Markov chain on $\Omega \subseteq \mathbb{R}^d$
- Precise definition of Markov chains on $\mathbb{R}^d$ is beyond the scope of this course. We will just state the most important results without proof. Assume for simplicity that $p(x) > 0$ for all $x \in \Omega$

Metropolis-Hastings algorithm on $\mathbb{R}^d$

- The Markov chain $X_0, X_1, \ldots$ on $\Omega \subseteq \mathbb{R}^d$ is irreducible if for any $x \in \Omega$ and $A \subset \Omega$, there is n such that

$$\mathbb{P}(X_n \in A | X_0 = x) > 0$$

Theorem

If the Metropolis-Hastings chain is irreducible, then for any function ϕ such that $\mathbb{E}_p[|\phi(X)|] < \infty$, the MH estimator is strongly consistent

$$\widehat{\theta}_n^{MH} = \frac{1}{n} \sum_{t=0}^{n-1} \phi(X_t) \to \theta \quad \textit{almost surely as } n \to \infty$$

Example: Gaussian distribution

- Let $X \sim N(\mu, \sigma^2)$ with

$$p(x) = (2\pi\sigma^2)^{-1/2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- MH algorithm with target pdf p and proposal transition pdf

$$q(y|x) = \begin{cases} 1 & \text{for } y \in [x - 1/2, x + 1/2] \\ 0 & \text{otherwise} \end{cases}$$

- Acceptance probability

$$\alpha(y|x) = \min\left(1, \frac{p(y)q(x|y)}{p(x)q(y|x)}\right) = \min\left(1, e^{-\frac{(y-\mu)^2}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^2}}\right)$$

Example: Gaussian distribution

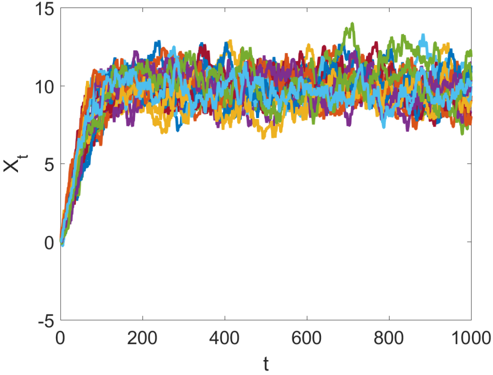


Figure: Realizations from the MH Markov chain $(X_0, \dots, X_{1000})$ with $X_0 = 0$.

Example: Gaussian distribution

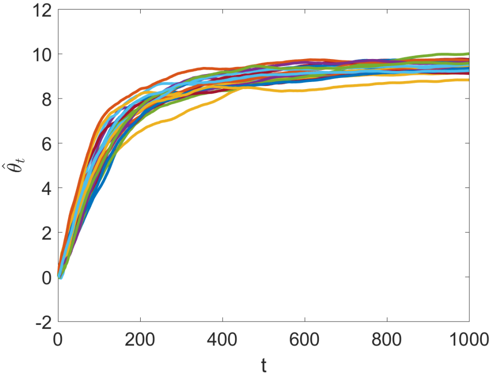
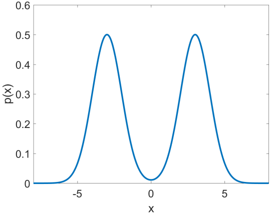


Figure: MH estimates $\widehat{\theta}_t = \frac{1}{t}\sum_{i=0}^{t-1} X_i$ of $\theta = \mathbb{E}_p[X]$ for different realizations of the Markov chain.

Example: Mixture of Gaussians distribution

- Let $p(x) = \pi_1(2\pi\sigma_1^2)^{-1/2}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \pi_2(2\pi\sigma_2^2)^{-1/2}e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$



- MH algorithm with target pdf p and proposal transition pdf

$$q(y|x) = \begin{cases} 1 & \text{for } y \in [x - 1/2, x + 1/2] \\ 0 & \text{otherwise} \end{cases}$$

- Acceptance probability

$$\alpha(y|x) = \min\left(1, \frac{p(y)q(x|y)}{p(x)q(y|x)}\right) = \min\left(1, p(y)/p(x)\right)$$

Example: Mixture of Gaussians distribution

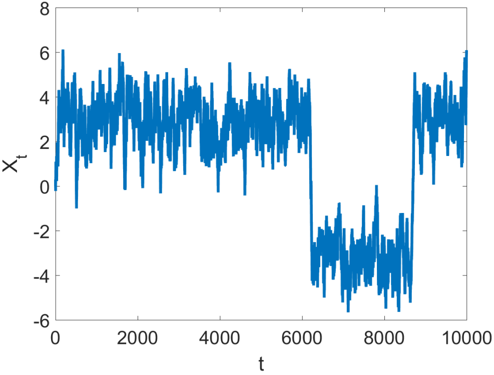


Figure: Realization of the Markov chain $(X_0, \ldots, X_{10000})$ with $X_0 = 0$.

Example

Consider a Metropolis algorithm for simulating samples from a bivariate Normal distribution with mean $\mu = (0, 0)$ and covariance

$$\Sigma = \begin{pmatrix} 1 & 0.7\sqrt{2} \\ 0.7\sqrt{2} & 2 \end{pmatrix}$$

using a proposal distribution for each component

$$x_i' | x_i \sim U(x - w, x + w) \quad i \in \{1, 2\}$$

The performance of the algorithm can be controlled by setting w.

- If w is small then we propose only small moves and the chain will move slowly around the parameter space.
- If w is large then it is large then we may only accept a few moves.

There is an 'art' to implementing MCMC that involves choice of good proposal distributions.

Example

Consider a Metropolis algorithm for simulating samples from a bivariate Normal distribution with mean $\mu = (0,0)$ and covariance

$$\Sigma = \begin{pmatrix} 1 & 0.7\sqrt{2} \\ 0.7\sqrt{2} & 2 \end{pmatrix}$$

using a proposal distribution for each component

$$x_i' | x_i \sim U(x - w, x + w) \quad i \in \{1, 2\}$$

The performance of the algorithm can be controlled by setting w.

- If w is small then we propose only small moves and the chain will move slowly around the parameter space.
- If w is large then it is large then we may only accept a few moves.

There is an 'art' to implementing MCMC that involves choice of good proposal distributions.

Example

Consider a Metropolis algorithm for simulating samples from a bivariate Normal distribution with mean $\mu = (0, 0)$ and covariance

$$\Sigma = \begin{pmatrix} 1 & 0.7\sqrt{2} \\ 0.7\sqrt{2} & 2 \end{pmatrix}$$

using a proposal distribution for each component

$$x_i' | x_i \sim U(x - w, x + w) \quad i \in \{1, 2\}$$

The performance of the algorithm can be controlled by setting w.

- If w is small then we propose only small moves and the chain will move slowly around the parameter space.
- If w is large then it is large then we may only accept a few moves.

There is an ‘art’ to implementing MCMC that involves choice of good proposal distributions.

Example

Consider a Metropolis algorithm for simulating samples from a bivariate Normal distribution with mean $\mu = (0, 0)$ and covariance

$$\Sigma = \begin{pmatrix} 1 & 0.7\sqrt{2} \\ 0.7\sqrt{2} & 2 \end{pmatrix}$$

using a proposal distribution for each component

$$x_i'|x_i \sim U(x - w, x + w) \quad i \in \{1, 2\}$$

The performance of the algorithm can be controlled by setting w.

- If w is small then we propose only small moves and the chain will move slowly around the parameter space.
- If w is large then it is large then we may only accept a few moves.

There is an 'art' to implementing MCMC that involves choice of good proposal distributions.

Example

Consider a Metropolis algorithm for simulating samples from a bivariate Normal distribution with mean $\mu = (0, 0)$ and covariance

$$\Sigma = \begin{pmatrix} 1 & 0.7\sqrt{2} \\ 0.7\sqrt{2} & 2 \end{pmatrix}$$

using a proposal distribution for each component

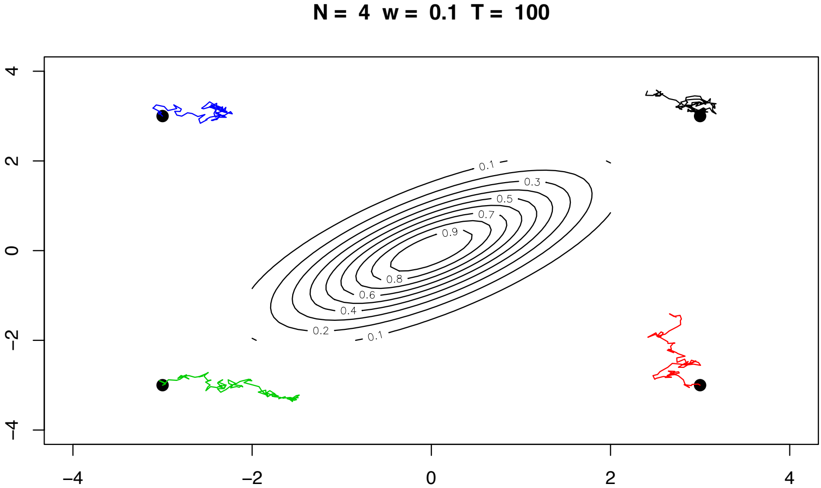
$$x_i' | x_i \sim U(x - w, x + w) \quad i \in \{1, 2\}$$

The performance of the algorithm can be controlled by setting w.

- If w is small then we propose only small moves and the chain will move slowly around the parameter space.
- If w is large then it is large then we may only accept a few moves.

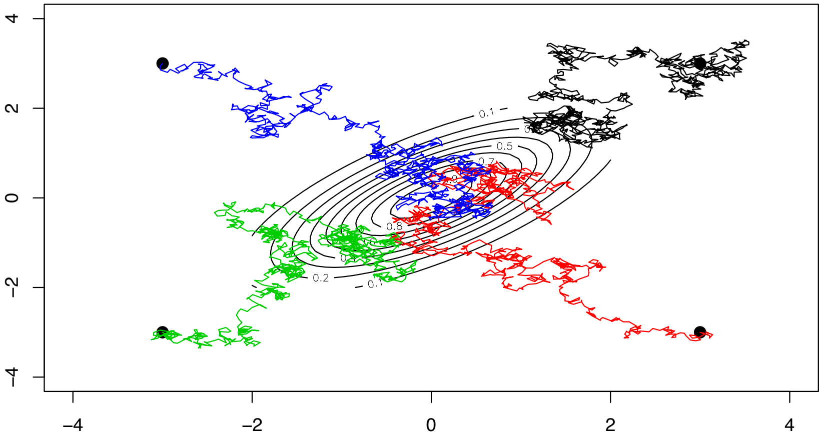
There is an ‘art’ to implementing MCMC that involves choice of good proposal distributions.

Example

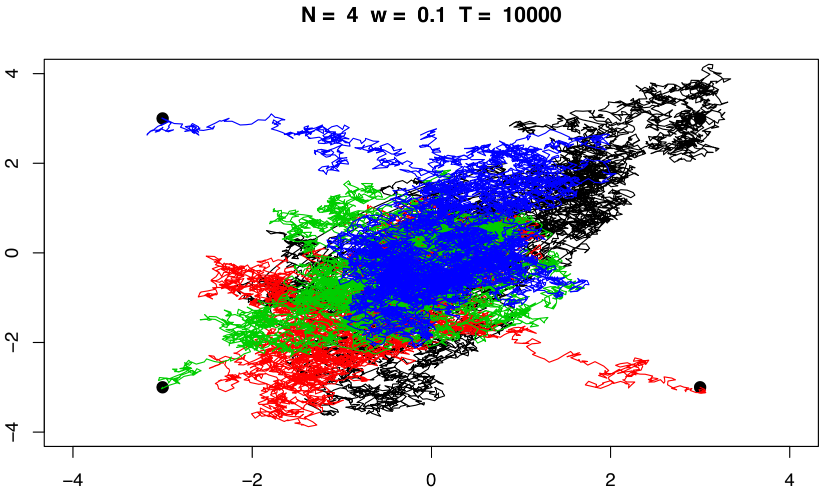


Example

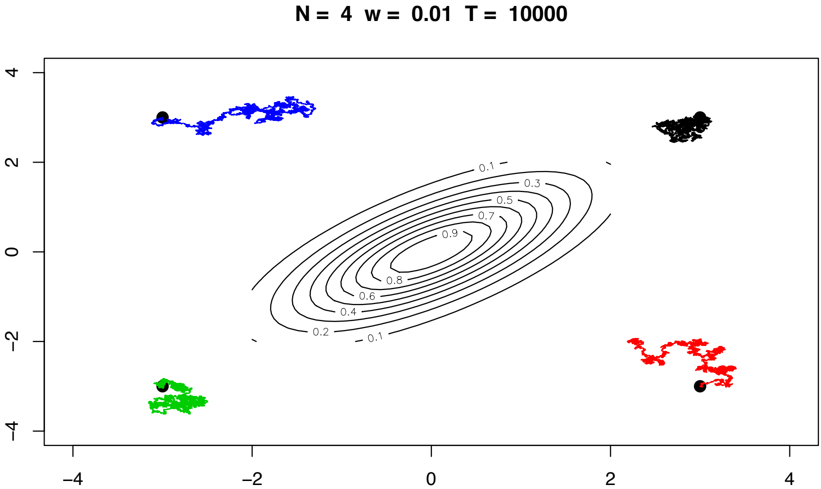
N = 4 w = 0.1 T = 1000



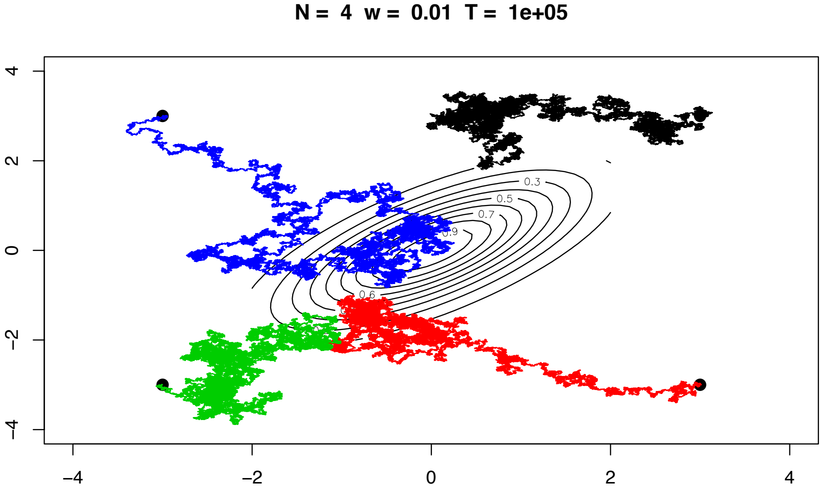
Example



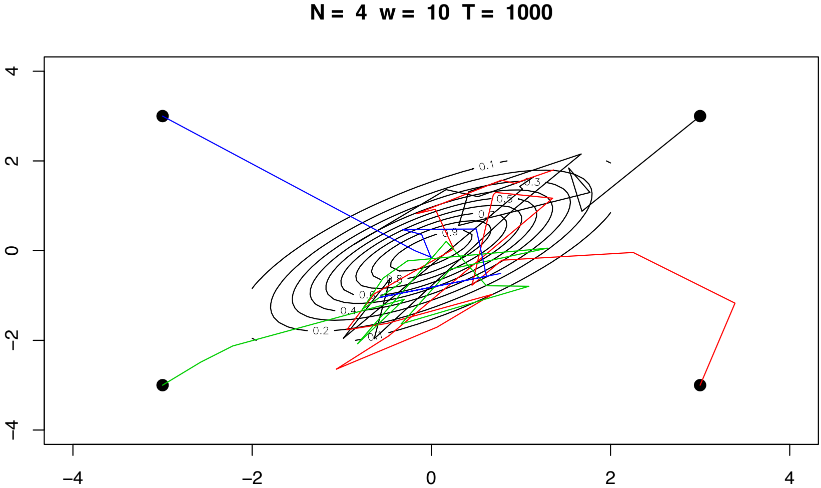
Example



Example



Example



MCMC: Practical aspects

- The MCMC chain does not start from the invariant distribution, so $\mathbb{E}[\phi(X_t)] \neq \mathbb{E}[\phi(X)]$ and the difference can be significant for small t
- X_t converges in distribution to p as $t \to \infty$
- Common practice is to discard the b first values of the Markov chain $X_0, \dots, X_{b-1}$, where we assume that X_b is approximately distributed from p
- We use the estimator
$$\frac{1}{n-b} \sum_{t=b}^{n-1} \phi(X_t)$$
- The initial $X_0, \dots, X_{b-1}$ is called the burn-in period of the MCMC chain.

MCMC: Gibbs Sampler

A Gibbs sampler is a particular type of MCMC algorithm that has been found to be very useful in high dimensional problems.

Suppose we wish to sample from $\pi(\theta)$ where $\theta = (\theta_1, \ldots, \theta_d)$. Each iteration of the Gibbs sampler occurs as follows

1. An ordering of the d components of θ is chosen.
2. For each component in this ordering, θ_j say, we draw a new value sampled from the full conditional distribution given all the other components of θ.

$$\theta_j^t \sim \pi(\theta_j | \theta_{-j}^{t-1})$$

where θ_{-j}^{t-1} represents all the components of θ, except the for θ_j, at their current values

$$\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1}).$$

MCMC: Gibbs Sampler

A Gibbs sampler is a particular type of MCMC algorithm that has been found to be very useful in high dimensional problems.

Suppose we wish to sample from $\pi(\theta)$ where $\theta = (\theta_1, \ldots, \theta_d)$. Each iteration of the Gibbs sampler occurs as follows

1. An ordering of the d components of θ is chosen.
2. For each component in this ordering, θ_j say, we draw a new value sampled from the full conditional distribution given all the other components of θ.

$$\theta_j^t \sim \pi(\theta_j | \theta_{-j}^{t-1})$$

where θ_{-j}^{t-1} represents all the components of θ, except the for θ_j, at their current values

$$\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1}).$$

MCMC: Gibbs Sampler

A Gibbs sampler is a particular type of MCMC algorithm that has been found to be very useful in high dimensional problems.

Suppose we wish to sample from $\pi(\theta)$ where $\theta = (\theta_1, \dots, \theta_d)$. Each iteration of the Gibbs sampler occurs as follows

1. An ordering of the d components of θ is chosen.
2. For each component in this ordering, θ_j say, we draw a new value sampled from the full conditional distribution given all the other components of θ.

$$\theta_j^t \sim \pi(\theta_j | \theta_{-j}^{t-1})$$

where θ_{-j}^{t-1} represents all the components of θ, except the for θ_j, at their current values

$$\theta_{-j}^{t-1} = (\theta_1^t, \dots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \dots, \theta_d^{t-1}).$$

MCMC: Gibbs Sampler

A Gibbs sampler is a particular type of MCMC algorithm that has been found to be very useful in high dimensional problems.

Suppose we wish to sample from $\pi(\theta)$ where $\theta = (\theta_1, \ldots, \theta_d)$. Each iteration of the Gibbs sampler occurs as follows

1. An ordering of the d components of θ is chosen.
2. For each component in this ordering, θ_j say, we draw a new value sampled from the full conditional distribution given all the other components of θ.

$$\theta_j^t \sim \pi(\theta_j | \theta_{-j}^{t-1})$$

where θ_{-j}^{t-1} represents all the components of θ, except the for θ_j, at their current values

$$\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1}).$$

MCMC: Gibbs Sampler

A Gibbs sampler is a particular type of MCMC algorithm that has been found to be very useful in high dimensional problems.

Suppose we wish to sample from $\pi(\theta)$ where $\theta = (\theta_1, \ldots, \theta_d)$. Each iteration of the Gibbs sampler occurs as follows

1. An ordering of the d components of θ is chosen.
2. For each component in this ordering, θ_j say, we draw a new value sampled from the full conditional distribution given all the other components of θ.

$$\theta_j^t \sim \pi(\theta_j | \theta_{-j}^{t-1})$$

where θ_{-j}^{t-1} represents all the components of θ, except the for θ_j, at their current values

$$\theta_{-j}^{t-1} = (\theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1}).$$

Example

Consider a single observation $y = (y_1, y_2)$ from a bivariate Normal distribution with unknown mean $\theta = (\theta_1, \theta_2)$ and known covariance $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. With a uniform prior on θ, the posterior distribution is

$$\theta | y \sim \mathsf{N}(y, \Sigma)$$

In this case the posterior is tractable but we will consider how to construct a Gibbs sampler.

We need the two conditional distributions $\pi(\theta_1 | \theta_2, y)$ and $\pi(\theta_2 | \theta_1, y)$.

Example

Consider a single observation $y = (y_1, y_2)$ from a bivariate Normal distribution with unknown mean $\theta = (\theta_1, \theta_2)$ and known covariance $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. With a uniform prior on θ, the posterior distribution is

$$\theta | y \sim \mathsf{N}(y, \Sigma)$$

In this case the posterior is tractable but we will consider how to construct a Gibbs sampler.

We need the two conditional distributions $\pi(\theta_1 | \theta_2, y)$ and $\pi(\theta_2 | \theta_1, y)$.

Bayesian statistics

Suppose that you observe y with pmf/pdf $f(y|\theta)$ where y and $\theta = (\theta_1, \ldots, \theta_p)$ can be large dimensional. θ is a parameter of the model that you are trying to estimate,

In Bayesian statistics we treat θ as a random variable. We choose $\pi(\theta)$ a *prior* distribution that represents our belief about the value of the parameter θ.

Bayes' Theorem tells us that once we observe y we should update our belief

$$\pi(\theta|y) = [f(y|\theta)\pi(\theta)]/f(y) \propto f(y|\theta)\pi(\theta).$$

Bayesian statistics

Suppose that you observe y with pmf/pdf $f(y|\theta)$ where y and $\theta = (\theta_1, \dots, \theta_p)$ can be large dimensional. θ is a parameter of the model that you are trying to estimate,
In Bayesian statistics we treat θ as a random variable. We choose $\pi(\theta)$ a *prior* distribution that represents our belief about the value of the parameter θ.
Bayes' Theorem tells us that once we observe y we should update our belief

$$\pi(\theta|y) = [f(y|\theta)\pi(\theta)]/f(y) \propto f(y|\theta)\pi(\theta).$$

Bayesian statistics

Suppose that you observe y with pmf/pdf $f(y|\theta)$ where y and $\theta = (\theta_1, \ldots, \theta_p)$ can be large dimensional. θ is a parameter of the model that you are trying to estimate,

In Bayesian statistics we treat θ as a random variable. We choose $\pi(\theta)$ a *prior* distribution that represents our belief about the value of the parameter θ.

Bayes' Theorem tells us that once we observe y we should update our belief

$$\pi(\theta|y) = [f(y|\theta)\pi(\theta)]/f(y) \propto f(y|\theta)\pi(\theta).$$

Example

$$\begin{aligned}
\pi(\theta_1|\theta_2, y) &\propto \pi(\theta_1, \theta_2|y) \\
&\propto \exp\left\{ -\frac{1}{2}(\theta - y)^T \Sigma^{-1} (\theta - y) \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix}^T \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix} \right\} \\
&= \exp\left\{ (\theta_1 - y_1)^2 + 2\rho(\theta_1 - y_1)(\theta_2 - y_2) + (\theta_2 - y_2)^2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} (\theta_1 - (y_1 + \rho(\theta_2 - y_2)))^2 \right\}
\end{aligned}$$

$$\Rightarrow \theta_1|\theta_2, y \sim \mathsf{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

Similarly,

$$\theta_2|\theta_1, y \sim \mathsf{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

Example

$$\begin{aligned}
\pi(\theta_1|\theta_2, y) &\propto \pi(\theta_1, \theta_2|y) \\
&\propto \exp\left\{ -\frac{1}{2}(\theta - y)^T \Sigma^{-1} (\theta - y) \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix}^T \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix} \right\} \\
&= \exp\left\{ (\theta_1 - y_1)^2 + 2\rho(\theta_1 - y_1)(\theta_2 - y_2) + (\theta_2 - y_2)^2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} (\theta_1 - (y_1 + \rho(\theta_2 - y_2)))^2 \right\}
\end{aligned}$$

$$\Rightarrow \theta_1|\theta_2, y \sim \mathsf{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

Similarly,

$$\theta_2|\theta_1, y \sim \mathsf{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

Example

$$\begin{aligned}
\pi(\theta_1|\theta_2,y) &\propto \pi(\theta_1,\theta_2|y) \\
&\propto \exp\left\{-\frac{1}{2}(\theta-y)^T\Sigma^{-1}(\theta-y)\right\} \\
&\propto \exp\left\{-\frac{1}{2(1-\rho^2)}\begin{pmatrix}\theta_1-y_1\\ \theta_2-y_2\end{pmatrix}^T\begin{pmatrix}1 & -\rho\\ -\rho & 1\end{pmatrix}\begin{pmatrix}\theta_1-y_1\\ \theta_2-y_2\end{pmatrix}\right\} \\
&= \exp\left\{(\theta_1-y_1)^2+2\rho(\theta_1-y_1)(\theta_2-y_2)+(\theta_2-y_2)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2(1-\rho^2)}(\theta_1-(y_1+\rho(\theta_2-y_2)))^2\right\}
\end{aligned}$$

$$\Rightarrow \theta_1|\theta_2,y \sim \mathsf{N}(y_1+\rho(\theta_2-y_2),1-\rho^2)$$

Similarly,

$$\theta_2|\theta_1,y \sim \mathsf{N}(y_2+\rho(\theta_1-y_1),1-\rho^2)$$

Example

$$
\begin{aligned}
\pi(\theta_1|\theta_2, y) &\propto \pi(\theta_1, \theta_2|y) \\
&\propto \exp\left\{ -\frac{1}{2}(\theta - y)^T \Sigma^{-1} (\theta - y) \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix}^T \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix} \right\} \\
&= \exp\left\{ (\theta_1 - y_1)^2 + 2\rho(\theta_1 - y_1)(\theta_2 - y_2) + (\theta_2 - y_2)^2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} (\theta_1 - (y_1 + \rho(\theta_2 - y_2)))^2 \right\}
\end{aligned}
$$

$$\Rightarrow \theta_1|\theta_2, y \sim \mathsf{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

Similarly,

$$\theta_2|\theta_1, y \sim \mathsf{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

Example

$$
\begin{aligned}
\pi(\theta_1|\theta_2, y) &\propto \pi(\theta_1, \theta_2|y) \\
&\propto \exp\left\{ -\frac{1}{2}(\theta - y)^T \Sigma^{-1} (\theta - y) \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix}^T \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix} \right\} \\
&= \exp\left\{ (\theta_1 - y_1)^2 + 2\rho(\theta_1 - y_1)(\theta_2 - y_2) + (\theta_2 - y_2)^2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} (\theta_1 - (y_1 + \rho(\theta_2 - y_2)))^2 \right\}
\end{aligned}
$$

$$\Rightarrow \theta_1|\theta_2, y \sim \mathsf{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

Similarly,

$$\theta_2|\theta_1, y \sim \mathsf{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

Example

$$\begin{aligned}
\pi(\theta_1|\theta_2, y) &\propto \pi(\theta_1, \theta_2|y) \\
&\propto \exp\left\{ -\frac{1}{2}(\theta - y)^T \Sigma^{-1} (\theta - y) \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix}^T \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix} \right\} \\
&= \exp\left\{ (\theta_1 - y_1)^2 + 2\rho(\theta_1 - y_1)(\theta_2 - y_2) + (\theta_2 - y_2)^2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} (\theta_1 - (y_1 + \rho(\theta_2 - y_2)))^2 \right\}
\end{aligned}$$

$$\Rightarrow \theta_1|\theta_2, y \sim \mathsf{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

Similarly,

$$\theta_2|\theta_1, y \sim \mathsf{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

Example

$$
\begin{aligned}
\pi(\theta_1|\theta_2, y) &\propto \pi(\theta_1, \theta_2|y) \\
&\propto \exp\left\{ -\frac{1}{2}(\theta - y)^T \Sigma^{-1} (\theta - y) \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix}^T \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix} \right\} \\
&= \exp\left\{ (\theta_1 - y_1)^2 + 2\rho(\theta_1 - y_1)(\theta_2 - y_2) + (\theta_2 - y_2)^2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} (\theta_1 - (y_1 + \rho(\theta_2 - y_2)))^2 \right\}
\end{aligned}
$$

$$\Rightarrow \theta_1|\theta_2, y \sim \mathsf{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

Similarly,

$$\theta_2|\theta_1, y \sim \mathsf{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

Example

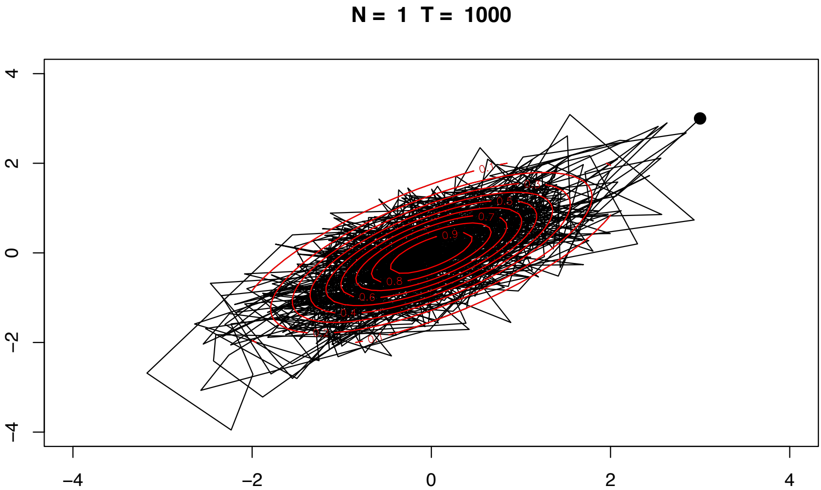
$$\begin{aligned}
\pi(\theta_1|\theta_2, y) &\propto \pi(\theta_1, \theta_2|y) \\
&\propto \exp\left\{ -\frac{1}{2}(\theta - y)^T \Sigma^{-1} (\theta - y) \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix}^T \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} \theta_1 - y_1 \\ \theta_2 - y_2 \end{pmatrix} \right\} \\
&= \exp\left\{ (\theta_1 - y_1)^2 + 2\rho(\theta_1 - y_1)(\theta_2 - y_2) + (\theta_2 - y_2)^2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2(1-\rho^2)} (\theta_1 - (y_1 + \rho(\theta_2 - y_2)))^2 \right\}
\end{aligned}$$

$$\Rightarrow \theta_1|\theta_2, y \sim \mathsf{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

Similarly,

$$\theta_2|\theta_1, y \sim \mathsf{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

Example



Deriving full conditional distributions

The Gibbs sampler requires us to be able to obtain the full conditional distributions of all components of the parameter vector θ.

To determine $\pi(\theta_j|\theta_{-j}^{t-1})$ we write down the posterior distribution and consider it as a function of θ_j.

If we are lucky then we will be able to ‘spot’ the distribution for $\theta_j|\theta_{-j}^{t-1}$ (using conjugate priors often helps).

If the full conditional is not of a nice form then this component could be sampled using rejection sampling or a MH update could be used instead of a Gibbs update.

Deriving full conditional distributions

The Gibbs sampler requires us to be able to obtain the full conditional distributions of all components of the parameter vector θ.

To determine $\pi(\theta_j|\theta_{-j}^{t-1})$ we write down the posterior distribution and consider it as a function of θ_j.

If we are lucky then we will be able to 'spot' the distribution for $\theta_j|\theta_{-j}^{t-1}$ (using conjugate priors often helps).

If the full conditional is not of a nice form then this component could be sampled using rejection sampling or a MH update could be used instead of a Gibbs update.

Deriving full conditional distributions

The Gibbs sampler requires us to be able to obtain the full conditional distributions of all components of the parameter vector θ.

To determine $\pi(\theta_j|\theta_{-j}^{t-1})$ we write down the posterior distribution and consider it as a function of θ_j.

If we are lucky then we will be able to 'spot' the distribution for $\theta_j|\theta_{-j}^{t-1}$ (using conjugate priors often helps).

If the full conditional is not of a nice form then this component could be sampled using rejection sampling or a MH update could be used instead of a Gibbs update.

Deriving full conditional distributions

The Gibbs sampler requires us to be able to obtain the full conditional distributions of all components of the parameter vector θ.

To determine $\pi(\theta_j|\theta_{-j}^{t-1})$ we write down the posterior distribution and consider it as a function of θ_j.

If we are lucky then we will be able to 'spot' the distribution for $\theta_j|\theta_{-j}^{t-1}$ (using conjugate priors often helps).

If the full conditional is not of a nice form then this component could be sampled using rejection sampling or a MH update could be used instead of a Gibbs update.

Gibbs sampling and Metropolis-Hastings

Gibbs sampling can be viewed as a special case of the MH algorithm with proposal distributions equal to the full conditionals.

The MH acceptance ratio is $\min(1, r)$ where

$$r = \frac{\pi(\theta')/q(\theta'|\theta^{t-1})}{\pi(\theta^{t-1})/q(\theta^{t-1}|\theta')}$$

We have $\theta^{t-1} = (\theta_1, \ldots, \theta_j, \ldots, \theta_d) = (\theta_j, \theta_{-j}^{t-1})$ and we use proposal

$$q(\theta'|\theta^{t-1}) = \pi(\theta_j'|\theta_{-j}^{t-1})$$

so that $\theta' = (\theta_1, \ldots, \theta_j', \ldots, \theta_d) = (\theta_j', \theta_{-j}^{t-1})$ and

$$q(\theta^{t-1}|\theta') = \pi(\theta_j|\theta_{-j}^{t-1})$$

.

Gibbs sampling and Metropolis-Hastings

Gibbs sampling can be viewed as a special case of the MH algorithm with proposal distributions equal to the full conditionals.

The MH acceptance ratio is $\min(1, r)$ where

$$r = \frac{\pi(\theta')/q(\theta'|\theta^{t-1})}{\pi(\theta^{t-1})/q(\theta^{t-1}|\theta')}$$

We have $\theta^{t-1} = (\theta_1, \ldots, \theta_j, \ldots, \theta_d) = (\theta_j, \theta_{-j}^{t-1})$ and we use proposal

$$q(\theta'|\theta^{t-1}) = \pi(\theta_j'|\theta_{-j}^{t-1})$$

so that $\theta' = (\theta_1, \ldots, \theta_j', \ldots, \theta_d) = (\theta_j', \theta_{-j}^{t-1})$ and

$$q(\theta^{t-1}|\theta') = \pi(\theta_j|\theta_{-j}^{t-1})$$

.

Gibbs sampling and Metropolis-Hastings

Gibbs sampling can be viewed as a special case of the MH algorithm with proposal distributions equal to the full conditionals.

The MH acceptance ratio is $\min(1, r)$ where

$$r = \frac{\pi(\theta')/q(\theta'|\theta^{t-1})}{\pi(\theta^{t-1})/q(\theta^{t-1}|\theta')}$$

We have $\theta^{t-1} = (\theta_1, \ldots, \theta_j, \ldots, \theta_d) = (\theta_j, \theta_{-j}^{t-1})$ and we use proposal

$$q(\theta'|\theta^{t-1}) = \pi(\theta_j'|\theta_{-j}^{t-1})$$

so that $\theta' = (\theta_1, \ldots, \theta_j', \ldots, \theta_d) = (\theta_j', \theta_{-j}^{t-1})$ and

$$q(\theta^{t-1}|\theta') = \pi(\theta_j|\theta_{-j}^{t-1})$$

.

Gibbs sampling and Metropolis-Hastings

Gibbs sampling can be viewed as a special case of the MH algorithm with proposal distributions equal to the full conditionals.

The MH acceptance ratio is $\min(1, r)$ where

$$r = \frac{\pi(\theta')/q(\theta'|\theta^{t-1})}{\pi(\theta^{t-1})/q(\theta^{t-1}|\theta')}$$

We have $\theta^{t-1} = (\theta_1, \ldots, \theta_j, \ldots, \theta_d) = (\theta_j, \theta_{-j}^{t-1})$ and we use proposal

$$q(\theta'|\theta^{t-1}) = \pi(\theta_j'|\theta_{-j}^{t-1})$$

so that $\theta' = (\theta_1, \ldots, \theta_j', \ldots, \theta_d) = (\theta_j', \theta_{-j}^{t-1})$ and

$$q(\theta^{t-1}|\theta') = \pi(\theta_j|\theta_{-j}^{t-1})$$

.

Gibbs sampling and Metropolis-Hastings

Gibbs sampling can be viewed as a special case of the MH algorithm with proposal distributions equal to the full conditionals.

The MH acceptance ratio is $\min(1, r)$ where

$$r = \frac{\pi(\theta')/q(\theta'|\theta^{t-1})}{\pi(\theta^{t-1})/q(\theta^{t-1}|\theta')}$$

We have $\theta^{t-1} = (\theta_1, \ldots, \theta_j, \ldots, \theta_d) = (\theta_j, \theta_{-j}^{t-1})$ and we use proposal

$$q(\theta'|\theta^{t-1}) = \pi(\theta_j'|\theta_{-j}^{t-1})$$

so that $\theta' = (\theta_1, \ldots, \theta_j', \ldots, \theta_d) = (\theta_j', \theta_{-j}^{t-1})$ and

$$q(\theta^{t-1}|\theta') = \pi(\theta_j|\theta_{-j}^{t-1})$$

.

Gibbs sampling and Metropolis-Hastings

Gibbs sampling can be viewed as a special case of the MH algorithm with proposal distributions equal to the full conditionals.

The MH acceptance ratio is $\min(1, r)$ where

$$r = \frac{\pi(\theta')/q(\theta'|\theta^{t-1})}{\pi(\theta^{t-1})/q(\theta^{t-1}|\theta')}$$

We have $\theta^{t-1} = (\theta_1, \ldots, \theta_j, \ldots, \theta_d) = (\theta_j, \theta_{-j}^{t-1})$ and we use proposal

$$q(\theta'|\theta^{t-1}) = \pi(\theta_j'|\theta_{-j}^{t-1})$$

so that $\theta' = (\theta_1, \ldots, \theta_j', \ldots, \theta_d) = (\theta_j', \theta_{-j}^{t-1})$ and

$$q(\theta^{t-1}|\theta') = \pi(\theta_j|\theta_{-j}^{t-1})$$

.

Gibbs sampling and Metropolis-Hastings

Then the MH acceptance ratio is $\min(1,r)$ where

$$r = \frac{\pi(\theta')/q(\theta'_j|\theta^{t-1})}{\pi(\theta^{t-1})/q(\theta_j^{t-1}|\theta')} = \frac{\pi(\theta')/\pi(\theta'_j|\theta_{-j}^{t-1})}{\pi(\theta^{t-1})/\pi(\theta_j|\theta_{-j}^{t-1})}$$

But the numerator can be simplified

$$\frac{\pi(\theta')}{\pi(\theta'_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta'_j, \theta_{-j}^{t-1})}{\pi(\theta'_j|\theta_{-j}^{t-1})} = \pi(\theta_{-j}^{t-1})$$

Similarly the denominator can be simplified

$$\frac{\pi(\theta^{t-1})}{\pi(\theta_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta_j, \theta_{-j}^{t-1})}{\pi(\theta_j|\theta_{-j}^{t-1})} = \pi(\theta_{-j}^{t-1})$$

So we have $r = \frac{\pi(\theta')/\pi(\theta'_j|\theta_{-j}^{t-1})}{\pi(\theta^{t-1})/\pi(\theta_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta_{-j}^{t-1})}{\pi(\theta_{-j}^{t-1})} = 1$

Thus, every proposed move is accepted.

Gibbs sampling and Metropolis-Hastings

Then the MH acceptance ratio is $\min(1, r)$ where

$$r = \frac{\pi(\theta')/q(\theta'_j|\theta^{t-1})}{\pi(\theta^{t-1})/q(\theta_j^{t-1}|\theta')} = \frac{\pi(\theta')/\pi(\theta'_j|\theta_{-j}^{t-1})}{\pi(\theta^{t-1})/\pi(\theta_j|\theta_{-j}^{t-1})}$$

But the numerator can be simplified

$$\frac{\pi(\theta')}{\pi(\theta'_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta'_j, \theta_{-j}^{t-1})}{\pi(\theta'_j|\theta_{-j}^{t-1})} = \pi(\theta_{-j}^{t-1})$$

Similarly the denominator can be simplified

$$\frac{\pi(\theta^{t-1})}{\pi(\theta_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta_j, \theta_{-j}^{t-1})}{\pi(\theta_j|\theta_{-j}^{t-1})} = \pi(\theta_{-j}^{t-1})$$

So we have $r = \frac{\pi(\theta')/\pi(\theta'_j|\theta_{-j}^{t-1})}{\pi(\theta^{t-1})/\pi(\theta_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta_{-j}^{t-1})}{\pi(\theta_{-j}^{t-1})} = 1$

Thus, every proposed move is accepted.

Gibbs sampling and Metropolis-Hastings

Then the MH acceptance ratio is $\min(1, r)$ where

$$r = \frac{\pi(\theta')/q(\theta'_j|\theta^{t-1})}{\pi(\theta^{t-1})/q(\theta_j^{t-1}|\theta')} = \frac{\pi(\theta')/\pi(\theta'_j|\theta_{-j}^{t-1})}{\pi(\theta^{t-1})/\pi(\theta_j|\theta_{-j}^{t-1})}$$

But the numerator can be simplified

$$\frac{\pi(\theta')}{\pi(\theta'_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta'_j, \theta_{-j}^{t-1})}{\pi(\theta'_j|\theta_{-j}^{t-1})} = \pi(\theta_{-j}^{t-1})$$

Similarly the denominator can be simplified

$$\frac{\pi(\theta^{t-1})}{\pi(\theta_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta_j, \theta_{-j}^{t-1})}{\pi(\theta_j|\theta_{-j}^{t-1})} = \pi(\theta_{-j}^{t-1})$$

So we have $r = \frac{\pi(\theta')/\pi(\theta'_j|\theta_{-j}^{t-1})}{\pi(\theta^{t-1})/\pi(\theta_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta_{-j}^{t-1})}{\pi(\theta_{-j}^{t-1})} = 1$

Thus, every proposed move is accepted.

Gibbs sampling and Metropolis-Hastings

Then the MH acceptance ratio is $\min(1, r)$ where

$$r = \frac{\pi(\theta')/q(\theta'_j|\theta^{t-1})}{\pi(\theta^{t-1})/q(\theta_j^{t-1}|\theta')} = \frac{\pi(\theta')/\pi(\theta'_j|\theta_{-j}^{t-1})}{\pi(\theta^{t-1})/\pi(\theta_j|\theta_{-j}^{t-1})}$$

But the numerator can be simplified

$$\frac{\pi(\theta')}{\pi(\theta'_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta'_j, \theta_{-j}^{t-1})}{\pi(\theta'_j|\theta_{-j}^{t-1})} = \pi(\theta_{-j}^{t-1})$$

Similarly the denominator can be simplified

$$\frac{\pi(\theta^{t-1})}{\pi(\theta_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta_j, \theta_{-j}^{t-1})}{\pi(\theta_j|\theta_{-j}^{t-1})} = \pi(\theta_{-j}^{t-1})$$

So we have $r = \frac{\pi(\theta')/\pi(\theta'_j|\theta_{-j}^{t-1})}{\pi(\theta^{t-1})/\pi(\theta_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta_{-j}^{t-1})}{\pi(\theta_{-j}^{t-1})} = 1$

Thus, every proposed move is accepted.

Gibbs sampling and Metropolis-Hastings

Then the MH acceptance ratio is $\min(1, r)$ where

$$r = \frac{\pi(\theta')/q(\theta'_j|\theta^{t-1})}{\pi(\theta^{t-1})/q(\theta_j^{t-1}|\theta')} = \frac{\pi(\theta')/\pi(\theta'_j|\theta_{-j}^{t-1})}{\pi(\theta^{t-1})/\pi(\theta_j|\theta_{-j}^{t-1})}$$

But the numerator can be simplified

$$\frac{\pi(\theta')}{\pi(\theta'_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta'_j, \theta_{-j}^{t-1})}{\pi(\theta'_j|\theta_{-j}^{t-1})} = \pi(\theta_{-j}^{t-1})$$

Similarly the denominator can be simplified

$$\frac{\pi(\theta^{t-1})}{\pi(\theta_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta_j, \theta_{-j}^{t-1})}{\pi(\theta_j|\theta_{-j}^{t-1})} = \pi(\theta_{-j}^{t-1})$$

So we have $r = \frac{\pi(\theta')/\pi(\theta'_j|\theta_{-j}^{t-1})}{\pi(\theta^{t-1})/\pi(\theta_j|\theta_{-j}^{t-1})} = \frac{\pi(\theta_{-j}^{t-1})}{\pi(\theta_{-j}^{t-1})} = 1$

Thus, every proposed move is accepted.

In its basic version, Gibbs sampling is a special case of the MetropolisHastings algorithm. However, in its extended versions (see below), it can be considered a general framework for sampling from a large set of variables by sampling each variable (or in some cases, each group of variables) in turn, and can incorporate the MetropolisHastings algorithm (or more sophisticated methods such as slice sampling, adaptive rejection sampling and adaptive rejection Metropolis algorithms) to implement one or more of the sampling steps.

Gibbs sampling is applicable when the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution of each variable is known and is easy (or at least, easier) to sample from. The Gibbs sampling algorithm generates an instance from the distribution of each variable in turn, conditional on the current values of the other variables. It can be shown that the sequence of samples constitutes a Markov chain, and the stationary distribution of that Markov chain is just the sought-after joint distribution.

Gibbs sampling is particularly well-adapted to sampling the posterior distribution of a Bayesian network, since Bayesian networks are typically specified as a collection of conditional distributions.