Chapter 4
Markov Chain Monte Carlo

Qing Zhou*

Contents

1 The Basic Idea .................................................. 2
  1.1 Markov chain Monte Carlo ................................. 2
  1.2 Transition kernel ........................................... 2
  1.3 Simulating a Markov chain ............................... 2
2 The Metropolis-Hastings Algorithm .......................... 3
  2.1 Algorithm .................................................... 3
  2.2 Detail balance and reversibility ........................... 5
  2.3 Temperature and simulated annealing ..................... 8
3 Convergence of MCMC ......................................... 9
  3.1 Limiting distribution and averages ....................... 9
  3.2 Autocorrelation and efficiency ........................... 9
4 Some Special Designs ......................................... 11
  4.1 Random-walk Metropolis ................................ 11
  4.2 Metropolized independence sampler (MIS) ............. 11
  4.3 Single-coordinate updating ................................ 13
5 The Gibbs Sampler .............................................. 15
  5.1 Random-scan Gibbs sampler .............................. 15
  5.2 Systematic-scan Gibbs sampler ........................... 15
  5.3 Stationary distribution .................................... 15
  5.4 Detail balance .............................................. 16
6 Examples of the Gibbs Sampler ............................... 19
  6.1 The slice sampler .......................................... 19
  6.2 Metropolized Gibbs sampler ............................ 20
  6.3 Blocked Gibbs sampler ................................... 20
7 Bayesian Missing Data Problems ............................ 22
  7.1 Introduction to Bayesian inference ....................... 22
  7.2 Data augmentation ........................................ 25

* UCLA Department of Statistics (email: zhou@stat.ucla.edu).
1. The Basic Idea

Want to generate $x = (x_1, x_2, \cdots, x_d) \sim \pi$ and compute

$$\mu = \mathbb{E}_\pi(h(X)) = \int h(x)\pi(x)dx.$$ 

1.1. Markov chain Monte Carlo

Generate a Markov chain $x_1, x_2, \cdots, x_n$ by simulating $x_t \sim p(\cdot| x_{t-1})$, where $x_t = (x_{1t}, \cdots, x_{dt})$, such that as $n \to \infty$,

1. $\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} h(x_t) \to \mu,$
2. $x_n \sim \pi.$

Note that $x_1, x_2, \cdots, x_n$ are correlated.

1.2. Transition kernel

Denote the one-step transition kernel of an M.C. by $K(x, y) = p_{X_t|X_{t-1}}(y|x)$. If

$$\int \pi(x)K(x, y)dx = \pi(y), \text{ for all } y$$

then $\pi(x)$ is a stationary distribution of the Markov chain with one-step transition kernel $K(x, y)$:

$$X_t \sim \pi \implies X_{t+1} \sim \pi.$$ 

1.3. Simulating a Markov chain

Given initial state $x_0$, transition kernel $K(x, y)$, want to simulate an M.C. with the transition kernel for $t = 1, 2, \cdots, n$.

For $t = 1, 2, \cdots, n$,

Draw $x_t \sim K(x_{t-1}, \bullet)$,

$t = t + 1.$
2. The Metropolis-Hastings Algorithm

2.1. Algorithm

Given a random initial state \( x^{(0)} \) and a proposal distribution \( q(x, y) \).

For \( t = 1, 2, \ldots, n \),

1. Draw \( y \) from the proposed distribution \( q(x^{(t-1)}, y) \);
2. Compute the MH ratio \( r(x^{(t-1)}, y) = \min \left[ 1, \frac{\pi(y)q(y, x^{(t-1)})}{\pi(x^{(t-1)})q(x^{(t-1)}, y)} \right] \);
3. Draw \( u \sim \text{Unif}(0, 1) \) and update
   \[
   x^{(t)} = \begin{cases} 
   y, & \text{if } u \leq r(x^{(t-1)}, y); \\
   x^{(t-1)}, & \text{otherwise}. 
   \end{cases}
   \]

First development: Metropolis et al. (1953) with \( q(x, y) = q(y, x) \) (symmetric proposal).

\[
r(x, y) = \min \left[ 1, \frac{\pi(y)}{\pi(x)} \right] = \begin{cases} 
1, & \text{if } \pi(y) \geq \pi(x); \\
\frac{\pi(y)}{\pi(x)}, & \text{if } \pi(y) < \pi(x).
\end{cases}
\]

\[
q(x, y) : \text{Unif} \left( x - \delta, x + \delta \right). \\
y \sim \text{Unif} \left( x^{(t)} - \delta, x^{(t)} + \delta \right).
\]

\[
r(x^{(t)}, y) = \min \left[ 1, \frac{\pi(y)q(y, x^{(t)})}{\pi(x^{(t)})q(x^{(t)}, y)} \right] = \min \left[ 1, \frac{\pi(y)}{\pi(x^{(t)})} \right]
\]

\[
q(y, x^{(t)}) = q(x^{(t)}, y) = \frac{1}{2\delta}.
\]

Example 1. Draw \( N(0, 1) \) by an MH algorithm with \( \text{Unif}(x - 1, x + 1) \) as the proposal.

# R code for this example

```r
n=10000;
X=numeric(n);
X[1]=0;
a=0;
```
for(t in 2:n)
{
    Y=runif(1,X[t-1]-1,X[t-1]+1);
    r=min(1,exp(-0.5*Y^2)/exp(-0.5*X[t-1]^2));
    u=runif(1,0,1);
    if(u<r){X[t]=Y;a=a+1}else{X[t]=X[t-1]};
}

a/n # acceptance rate
[1] 0.805

#use the last 5000 iterations (X[5001:n]) as our samples from N(0,1)

mean(X[5001:n])
[1] -0.04334007

sd(X[5001:n])
[1] 0.9988046

hist(X[5001:n])

Some remarks:
- If an MH algorithm is irreducible and aperiodic, then the induced M.C. converges to the stationary distribution $\pi(x)$.
- Burn-in period. Use the $N(0,1)$ example with different initial values $x^{(0)} = 5$ (red) vs $x^{(0)} = -5$ (blue).
2.2. Detail balance and reversibility

Want to verify $\int \pi(x)K(x,y)dx = \pi(y)$. ($\sum x \pi(x)K(x,y) = \pi(y)$ for discrete cases)

More strict condition (but easy to check): detail balance.

If $\pi(x)K(x,y) = \pi(y)K(y,x)$, then $\pi$ is a stationary distribution of the Markov chain with $K(x,y)$ as the one-step transition kernel.

Proof. From the detail balance condition,

$$\int \pi(x)K(x,y)dx = \int \pi(y)K(y,x)dx = \pi(y) \int K(y,x)dx = \pi(y).$$

Definition 1. Markov chains that satisfy detail balance are called reversible.

Theorem 1. The MH algorithm induces a transition kernel w.r.t. which $\pi(x)$ is stationary (invariant).
Proof. Want to show $\pi(x)K(x, y) = \pi(y)K(y, x)$ for any $x$ and $y$.

1. It is trivial for $x = y$;
2. Suppose $y \neq x$. Then $K(x, y) = q(x, y) \min \left[ 1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \right]$ for the MH Algorithm. Now we have

\[
\pi(x)K(x, y) = \pi(x)q(x, y) \min \left[ 1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \right] \\
= \min[\pi(x)q(x, y), \pi(y)q(y, x)] \\
= \min \left[ \frac{\pi(x)q(x, y)}{\pi(y)q(y, x)}, 1 \right] \cdot \pi(y)q(y, x) \\
= \pi(y)K(y, x).
\]

Example 2. Poisson Distribution.

$\pi(x) = e^{-\lambda x} \frac{\lambda^x}{x!} \propto \frac{\lambda^x}{x!}$, for $x = 0, 1, 2, \ldots$.

Design proposal: $q(x, y) \begin{cases} 
\text{If } x \geq 1, & y = \begin{cases} 
x + 1, & \text{with probability } \frac{1}{2}; \\
x - 1, & \text{with probability } \frac{1}{2}; 
\end{cases} \\
\text{If } x = 0, & y = x + 1 \text{ with probability } 1.
\end{cases}$

The ratio between target densities: $\frac{\pi(y)}{\pi(x)} = \frac{\lambda^y y!}{\lambda^x x!}$ ($\pi(x)$ can be unnormalized.)

If $x, y \geq 1$, $\frac{q(y, x)}{q(x, y)} = 1$: Symmetric.

If $x = 0, y = 1$, $\frac{q(y, x)}{q(x, y)} = \frac{\frac{1}{2}}{1} = \frac{1}{2}$.

If $x = 1, y = 0$, $\frac{q(y, x)}{q(x, y)} = \frac{1}{\frac{1}{2}} = 2$. 
Example 3. 1-d Ising Model.

\[ \mathbf{x} = (x_1, \cdots, x_d), \quad U(x) = -\sum_{i=1}^{d-1} x_ix_{i+1}; \text{ energy.} \]

At temperature \( T \), Boltzmann distribution:

\[ \pi(x) \propto \exp \left( -\frac{U(x)}{T} \right) = \exp \left( \frac{1}{T} \sum_{i=1}^{d-1} x_ix_{i+1} \right). \]

Put \( \mu = 1/T > 0 \).

Current configuration \( x^{(t)} = (x_1^{(t)}, \cdots, x_d^{(t)}) \),

1. Randomly choose one spin \( j \), flip to the opposite, \( y = (x_1^{(t)}, \cdots, -x_j^{(t)}, \cdots, x_d^{(t)}) \);
2. Symmetric proposal: \( q(x^{(t)}, y) = q(y, x^{(t)}) \),

\[
\begin{align*}
r(x^{(t)}, y) & = \min \left[ 1, \frac{\pi(y)}{\pi(x^{(t)})} \right], \\
n(x^{(t)}, y) & = \exp \left\{ -2\mu x_j^{(t)} \left( x_{j-1}^{(t)} + x_{j+1}^{(t)} \right) \right\}.
\end{align*}
\]

#R code: Metropolis for Ising model, change the value of T to see its effect.

```r
n=6000;
d=20;
X=matrix(0,n,d);
X[1,]=sample(c(-1,1),size=d,replace=TRUE);
g=numeric(n);
g[1]=sum(X[1,]);
T=0.1;
for(t in 2:n)
{
    y=X[t-1,];
    j=sample(1:d,size=1);
    ```
\[ y[j] = -X[t-1,j]; \]
if \( j = 1 \) {
  \[ r = \exp(-2*X[t-1,1]*X[t-1,2]/T); \]
} else if \( j = d \) {
  \[ r = \exp(-2*X[t-1,d-1]*X[t-1,d]/T); \]
} else {
  \[ r = \exp(-2*X[t-1,j]*(X[t-1,j-1]+X[t-1,j+1])/T); \]
}
\[ U = \text{runif}(1,0,1); \]
if \( U \leq \min(r,1) \) \{ \[ X[t,\] = \]y \} else \{ \[ X[t,\] = X[t-1,\]; \}
\[ g[t] = \text{sum}(X[t,]); \]
\}
\[ \text{mean}(g[1000:n]) \]

### 2.3. Temperature and simulated annealing

For any \( \pi(x) \), let \( h(x) = -\log(\pi(x)) \). For \( T > 0 \), define

\[ \pi(x; T) \propto \exp \left( -\frac{h(x)}{T} \right). \]

In particular, \( \pi(x) = \pi(x; T = 1) \). Suppose \( x^\ast = \arg\min h(x) = \arg\max \pi(x) \).

- \( T \to \infty \): \( \pi(x; T) \propto 1 \), close to uniform distribution.
- \( T \to 0^+ \): \( \pi(x; T)/\pi(x^\ast; T) \to 0 \) for any \( h(x) > h(x^\ast) \). Thus \( \pi(x; T) \) is concentrated around \( x^\ast \).

**Simulated Annealing**: To find \( x^\ast \), the global minimizer of \( h(x) \).

Choose \( T_1 \geq T_2 \geq \cdots \geq T_n \to 0^+ \) and pick \( x^{(0)} \). For \( i = 1, \ldots, n \),

- Draw \( x^{(i)} \) given \( x^{(i-1)} \) via one step of an MH algorithm targeting at \( \pi(x; T_i) \).
3. Convergence of MCMC

3.1. Limiting distribution and averages

Goal of MCMC: to estimate

\[ \mu_h = \int h(x) \pi(x) dx = E_\pi[h(X)]. \]

Suppose \( x^{(1)}, x^{(2)}, \ldots, x^{(m)} \) is a Markov chain on finite discrete state space \( \mathcal{X} \) with one step transition kernel \( K(x, y) \). Let \( \overline{h}_m = \frac{1}{m} \sum_{i=1}^{m} h(x^{(i)}). \)

**Definition 2.** A state \( x \in \mathcal{X} \) is irreducible if under the transition kernel \( K \) there is positive probability to move from \( x \) to any other state and then coming back in a finite number of steps. A state \( x \) is aperiodic if the greatest common divider of \( \{n : K^{(n)}(x, x) > 0\} \) is 1, where \( K^{(n)} \) is the \( n \)-step transition kernel.

**Theorem 2.** Suppose a finite-state Markov chain \( x^{(0)}, x^{(1)}, \ldots \) is irreducible and aperiodic with \( \pi \) as a stationary distribution. Then for any \( x^{(0)} \),

1. \[ \| K^{(n)}(x^{(0)}, \cdot) - \pi \|_{\text{var}} = \frac{1}{2} \sum_{x \in \mathcal{X}} |K^{(n)}(x^{(0)}, x) - \pi(x)| \leq cr^n, \text{ for some } c > 0 \text{ and } r \in (0, 1); \]
2. \( \overline{h}_m \to \mu_h \) almost surely as \( m \to \infty \);
3. \( \sqrt{m}(\overline{h}_m - \mu_h) \to N(0, \sigma_h^2) \) in distribution as \( m \to \infty \).

3.2. Autocorrelation and efficiency

Assume that \( x^{(0)} \sim \pi(x) \). Then:

\[ m \text{ Var}(\overline{h}_m) = \sigma^2 \left[ 1 + 2 \sum_{j=1}^{m-1} \left( 1 - \frac{j}{m} \right) \rho_j \right] \approx \sigma^2 \left[ 1 + 2 \sum_{j=1}^{\infty} \rho_j \right] (m \to \infty), \]

where \( \sigma^2 = \text{Var}_\pi[h(x)], \rho_j = \text{cor}(h(x^{(1)}), h(x^{(j+1)})): j \)-step autocorrelation.

**Definition 3.** The integrated autocorrelation time: \( T_{int}(h) = \frac{1}{2} + \sum_{j=1}^{\infty} \rho_j. \)

Then \( m \cdot \text{Var}(\overline{h}_m) = 2T_{int}(h) \cdot \sigma^2 \Rightarrow \)

\[ \text{Var}(\overline{h}_m) = \frac{2T_{int}(h) \cdot \sigma^2}{m} = \frac{\sigma^2}{m/(2T_{int}(h))}. \]
Compare to independent sample $\text{Var}(\bar{h}_m) = \frac{\sigma^2}{m}$.

Effective sample size of this Markov chain is

$$\frac{m}{2T_{\text{int}}(h)} = \frac{m}{1 + 2\sum_{j=1}^{\infty} \rho_j}.$$
4. Some Special Designs.

4.1. Random-walk Metropolis

\( \pi(x) \) is defined on \( \mathbb{R}^d \) (d-dim Euclidean Space).

Use the addition of a random perturbation (error) as the proposal.

Proposal: \( y = x^{(t)} + \varepsilon_t, \quad \varepsilon_t \sim g_\sigma(\varepsilon) \): spherically symmetric, i.e., \( g_\sigma(a) = g_\sigma(b) \) if \( ||a|| = ||b|| \).

Examples: \( g_\sigma(\varepsilon) = N(0, \sigma^2 I_d) \) or \( g_\sigma(\varepsilon) = \text{Unif}(B(0, \sigma)) \).

\( \Rightarrow \) symmetric proposal \( q(x^{(t)}, y) = q(y, x^{(t)}): g_\sigma(\varepsilon) = g_\sigma(-\varepsilon) \).

The random-walk Metropolis:

Given \( x^{(t)} \),

1. Draw \( \varepsilon_t \sim g_\sigma(\varepsilon) \): spherically symmetric (\( \sigma \) can be controlled by the user),
   set \( y = x^{(t)} + \varepsilon_t, \ r(x^{(t)}, y) = \min \left[ 1, \frac{\pi(y)}{\pi(x^{(t)})} \right] \);

2. Draw \( u \sim \text{Unif}(0, 1) \) and update
   \[ x^{(t+1)} = \begin{cases} 
   y, & \text{if } u \leq r(x^{(t)}, y); \\
   x^{(t)}, & \text{otherwise.} 
   \end{cases} \]

How to choose \( \sigma \): maintain acceptance rate \( \in [25\%, 35\%] \).

4.2. Metropolized independence sampler (MIS)

\( q(x, y) \) is an independent trial density \( g(y) \), independent of \( x \).
Given \( x(t) \),

1. Draw \( y \sim g(y) \),
\[
r(x(t), y) = \min \left[ 1, \frac{\pi(y)}{\pi(x(t))} \frac{g(x(t))}{g(y)} \right] = \min \left[ 1, \frac{w(y)}{w(x(t))} \right],
\]
\( w(x) = \pi(x)/g(x) \) is the importance weight;
2. Draw \( u \sim \text{Unif}(0, 1) \),
\[
x(t+1) = \begin{cases} 
    y, & \text{if } u \leq r(x(t), y); \\
    x(t), & \text{otherwise}.
\end{cases}
\]

Remarks:

(a) Efficiency depends on how close \( g(y) \) is to \( \pi(y) \): one way to measure the closeness is by \( \text{Var}_g(w(x)) \). Small \( \text{Var}_g(w(x)) \) suggests that \( g \) is close to \( \pi \) and usually leads to a higher acceptance rate. If \( \text{Var}_g(w(x)) = 0 \), then \( g = \pi \) and \( r(x, y) = 1 \) for all \( x, y \).

(b) Robust performance: \( g \) should have a heavier tail than \( \pi \). For example, if \( \pi \) is normal then \( g \) could be a \( t \)-distribution.

\[
\text{Example 4 (Gamma distribution).} \text{ Design a Metropolized independent sampler to draw from Gamma}(\alpha, \beta), \alpha > 1, \beta > 0,
\]
\[
\pi(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0,
\]
using \( \text{Exp}(\lambda) \) as the trial distribution. Let
\[
w(x) = \frac{x^{\alpha-1} e^{-\beta x}}{\lambda e^{-\lambda x}}.
\]
Choose \( \lambda \) to minimize \( \text{Var}_g(w(x)) \).

Since \( \mathbb{E}_g(w(x)) = \int x^{\alpha-1} e^{-\beta x} dx = \Gamma(\alpha)/\beta^\alpha \) is a constant independent of \( \lambda \), it is equivalent to minimizing \( \mathbb{E}_g[w(x)^2] = \int w(x)^2 g(x) dx \). Some calculation shows that
\[
\mathbb{E}_g[w(x)^2] = \frac{1}{\lambda} \int_0^\infty x^{2\alpha-2} e^{-(2\beta - \lambda)x} dx.
\]
Note that $E_g[w(x)^2] < \infty$ if and only if $2\beta - \lambda > 0$. So we must choose
\[
\lambda < 2\beta. \tag{1}
\]
Under this condition, the integrand is an unnormalized Gamma$(2\alpha - 1, 2\beta - \lambda)$ and thus
\[
E_g[w(x)^2] = \frac{1}{\lambda} \cdot \frac{\Gamma(2\alpha - 1)}{(2\beta - \lambda)^{2\alpha - 1}}.
\]
Therefore, to minimize $E_g[w(x)^2]$ we just need to maximize
\[
f(\lambda) = \lambda(2\beta - \lambda)^{2\alpha - 1}
\]
over $\lambda$. Since the objective $f(\lambda) > 0$, we can equivalently
\[
\max_\lambda \left[ \log f(\lambda) = \log \lambda + (2\alpha - 1) \log(2\beta - \lambda) \right]
\]
of which the only maximizer is
\[
\lambda^* = \beta/\alpha
\]
by setting derivative to zero. Since $\alpha > 1$, we have $\lambda^* < \beta$ satisfying the constraint (1). This also shows that the tail of $g$ is heavier than that of $\pi$ (Remark b):
\[
\lim_{x \to \infty} \frac{\pi(x)}{g(x)} = C \lim_{x \to \infty} \frac{x^{\alpha - 1}}{e^{(\beta - \lambda^*)x}} = 0,
\]
where $C > 0$ is a constant.

In fact, with $\lambda^* = \beta/\alpha$, $g$ and $\pi$ have the same mean $(1/\lambda^* = \alpha/\beta)$. That is, we have matched the expectations of the two distributions with this optimal choice.

4.3. Single-coordinate updating

$x = (x_1, x_2, \ldots, x_i, \ldots, x_d)$.

$x_i(y) \triangleq (x_1, x_2, \ldots, y, \ldots, x_d)$: $x$ with $y$ replacing $x_i$.

$x_{[-i]} \triangleq (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$: $x$ with $x_i$ omitted.

$q(x, y)$ has two steps:

(a) Select a coordinate $i$ \{ deterministic, circular from 1 to $d$; random (uniformly). \}

(b) Draw $y \sim q_i(x_i, y)$ and put $y = x_i(y)$. 
The MH ratio is determined by
\[
\frac{\pi(y) q(y, x)}{\pi(x) q(x, y)} = \frac{\pi(x_i(y)) q_i(y, x_i)}{\pi(x) q_i(x_i, y)}
\]

Let \(\pi_i(\cdot | x_{[-i]})\) be the conditional density of \([x_i | x_{[-i]}]\). Then we have
\[
\pi(x) = \pi_i(x_i | x_{[-i]}) \cdot \pi(x_{[-i]}),
\]
\[
\pi(x_i(y)) = \pi_i(y | x_{[-i]}) \cdot \pi(x_{[-i]})
\]

\[\Rightarrow \frac{\pi(x_i(y))}{\pi(x)} = \frac{\pi_i(y | x_{[-i]})}{\pi_i(x_i | x_{[-i]})} \] ratio of conditional densities.

**Special case:** If we choose \(q_i(x_i, y) = \pi_i(y | x_{[-i]}) \& q_i(y, x_i) = \pi_i(x_i | x_{[-i]})\),

then \(r(x, y) = \min \left[1, \frac{\pi_i(y | x_{[-i]})}{\pi_i(x_i | x_{[-i]})} \cdot \frac{\pi_i(x_i | x_{[-i]})}{\pi_i(y | x_{[-i]})} \right] \equiv 1\), so \(y\) is always accepted.

\[\Rightarrow\] The Gibbs sampler.
5. The Gibbs Sampler

Target distribution: $\pi(x) = \pi(x_1, x_2, \cdots, x_d), x \in \mathbb{R}^d$. Put

\[
\begin{align*}
\mathbf{x}^{(t)} &= (x_1^{(t)}, x_2^{(t)}, \cdots, x_d^{(t)}), \\
x_i^{(t)}(y) &= (x_1^{(t)}, \cdots, x_{i-1}^{(t)}, y, x_{i+1}^{(t)}, \cdots, x_d^{(t)}), \\
x_{[-i]}^{(t)} &= (x_1^{(t)}, \cdots, x_{i-1}^{(t)}, x_{i+1}^{(t)}, \cdots, x_d^{(t)}).
\end{align*}
\]

5.1. Random-scan Gibbs sampler

Let $\mathbf{x}^{(t)}$ be the current sample for iteration $t$. Then at iteration $t+1$:

1. Randomly select a coordinate $i$ from $\{1, 2, \cdots, d\}$;
2. Draw $y$ from the conditional distribution $\pi(x_i | x_{[-i]})$. Let $\mathbf{x}^{(t+1)} = x_i^{(t)}(y)$ (i.e. $x_i^{(t+1)} = y, \mathbf{x}^{(t+1)}_{[-i]} = \mathbf{x}^{(t)}_{[-i]}$).

5.2. Systematic-scan Gibbs sampler

Let $\mathbf{x}^{(t)} = (x_1^{(t)}, \cdots, x_d^{(t)})$. At $t+1$ iteration:

For $i = 1, 2, \cdots, d$,
\[
draw \ x_i^{(t+1)} \sim \pi(x_i | x_1^{(t+1)}, \cdots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \cdots, x_d^{(t)}).
\]

5.3. Stationary distribution

$x = (x_1, \cdots, x_d), y = x_i(y)$, and transition kernel $K(x, y) = \pi(y | x_{[-i]})$.

\[
\begin{align*}
\int_{x_i} \pi(x)K(x, y)dx_i &= \int_{x_i} \pi(x_i, x_{[-i]})dx_i \pi(y | x_{[-i]}) \\
&= \pi(x_{[-i]})\pi(y | x_{[-i]}) \\
&= \pi(y, x_{[-i]}) = \pi(x_i(y)) \\
&= \pi(y).
\end{align*}
\]

Therefore, $\pi$ is a stationary distribution with respect to the transition kernel $K$. 

5.4. Detail balance

To verify detail balance, need to show that \( \pi(x)K(x,y) = \pi(y)K(y,x) \).

Proof.

\[
\pi(x)K(x,y) = \pi(x) \cdot \pi(y|x_{[-i]}) = \frac{\pi(x) \cdot \pi(y)}{\pi(x_{[-i]})}.
\]

\[
\pi(y)K(y,x) = \pi(y) \cdot \pi(x|x_{[-i]}) = \frac{\pi(x) \cdot \pi(y)}{\pi(x_{[-i]})}.
\]

Example 5. Bivariate Normal.

\( x = (x_1, x_2) \sim \mathcal{N}_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \), i.e.

\[
\pi(x_1, x_2) = \frac{1}{2\pi \sqrt{1-\rho^2}} \exp \left\{ -\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)} \right\}.
\]

Find the conditional distribution \( [x_1 \mid x_2] \) as follows: Regarding \( x_2 \) as a constant,

\[
\pi(x_1 \mid x_2) \propto \pi(x_1, x_2) \propto \exp \left[ -\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)} \right].
\]

Complete squares: \( x_1^2 - 2\rho x_1 x_2 = (x_1 - \rho x_2)^2 - (\rho x_2)^2 \). Plug into the above,

\[
\pi(x_1 \mid x_2) \propto \exp \left[ -\frac{(x_1 - \rho x_2)^2}{2(1-\rho^2)} \right],
\]

which is an unnormalized density for \( \mathcal{N}(\rho x_2, 1-\rho^2) \). Thus,

\[
x_1 \mid x_2 \sim \mathcal{N}(\rho x_2, 1 - \rho^2).
\]

Similarly, \( x_2 \mid x_1 \sim \mathcal{N}(\rho x_1, 1 - \rho^2) \).

Systematic-scan: Given \( x^{(t)} = (x_1^{(t)}, x_2^{(t)}) \),

\[
x_1^{(t+1)} \mid x_2^{(t)} \sim \mathcal{N}(\rho x_2^{(t)}, 1 - \rho^2). \tag{2}
\]

\[
x_2^{(t+1)} \mid x_1^{(t+1)} \sim \mathcal{N}(\rho x_1^{(t+1)}, 1 - \rho^2). \tag{3}
\]
#R code: Gibbs sampler for Example 5 (bivariate normal)

```r
rho = 0.8;
n = 5000;
X = matrix(0, n, 2);
X[1,] = c(10, 10);
for (t in 2:n) {
  X[t,1] = rnorm(1, rho*X[t-1,2], sqrt(1-rho^2));
  X[t,2] = rnorm(1, rho*X[t,1], sqrt(1-rho^2));
}
```

Use induction:

\[
\begin{pmatrix}
  x_1^{(t)} \\
  x_2^{(t)}
\end{pmatrix}
\sim \mathcal{N}_2\left(\begin{pmatrix} x_2^{(0)} \\ \rho x_2^{(0)} \end{pmatrix}, \begin{pmatrix} 1 - \rho^4 t - 2 & \rho - \rho^4 t - 1 \\ \rho - \rho^4 t - 1 & 1 - \rho^4 t \end{pmatrix}\right)
\]

\[
\xrightarrow{t \to \infty} \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).
\]

**Example 6.** \(\pi(x, y) \propto \binom{n}{x} y^{x + \alpha - 1}(1 - y)^{n - x + \beta - 1}\), for \(x = 0, 1, \ldots, n\) and \(y \in [0, 1]\).

\[
\pi(x|y) \propto \binom{n}{x} y^{x}(1 - y)^{n - x} \Rightarrow x|y \sim \text{Bin}(n, y).
\]

\[
\pi(y|x) \propto y^{x + \alpha - 1}(1 - y)^{n - x + \beta - 1} \Rightarrow y|x \sim \text{Beta}(x + \alpha, n - x + \beta).
\]
Beta Distribution: Density \( f(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \) for \( x \in [0, 1] \), \( \alpha > 0, \beta > 0 \).

\[ E(X) = \frac{\alpha}{\alpha + \beta}. \]

If \( X_1 \sim \text{Gamma}(\alpha, 1) \), \( X_2 \sim \text{Gamma}(\beta, 1) \), then \( \frac{X_1}{X_1 + X_2} \sim \text{Beta}(\alpha, \beta) \).

Example 7. 1-d Ising Model: For \( T > 0 \),

\[ \pi(x) \propto \exp \left( \frac{1}{T} \sum_{i=1}^{d-1} x_i x_{i+1} \right), \quad x_i \in \{1, -1\}. \]

Find the conditional distribution \([x_i \mid x_{\lnot i}]\).

\[ \pi(x_i \mid x_{\lnot i}) \propto \pi(x_1, \ldots, x_i, \ldots, x_d) \]

\[ \propto \exp \left\{ \frac{1}{T} \left( x_1 x_2 + \ldots + x_{i-1} x_i + x_i x_{i+1} + \ldots + x_{d-1} x_d \right) \right\} \]

\[ \propto \exp \left\{ \frac{x_i}{T} (x_{i-1} + x_{i+1}) \right\}. \]

Since \( x_i \in \{1, -1\} \), put \( Z_i = \exp \left\{ \frac{1}{T} (x_{i-1} + x_{i+1}) \right\} + \exp \left\{ -\frac{1}{T} (x_{i-1} + x_{i+1}) \right\} \).

We have

\[ \pi(x_i \mid x_{\lnot i}) = \frac{1}{Z_i} \exp \left\{ \frac{x_i}{T} (x_{i-1} + x_{i+1}) \right\} \quad \text{for } x_i \in \{1, -1\}. \]

For \( i = 1 \) or \( d \), plug in \( x_0 = x_{d+1} = 0 \).

Note that \( \pi(x_i \mid x_{\lnot i}) = P(X_i = x_i \mid x_{\lnot i}) \) (conditional probability).

Let \( \theta_i = \pi(x_i = 1 \mid x_{\lnot i}) \). To sample from \([x_i \mid x_{\lnot i}]\):

- Draw \( u \sim \text{Unif}(0, 1) \);
- Let \( x_i = 1 \) if \( u < \theta_i \) and \( x_i = -1 \) otherwise.
6. Examples of the Gibbs Sampler

6.1. The slice sampler

\( \pi(x) \): target density, \( x \in \mathbb{R}^d \).

**Lemma 1.** If \( (X, Y) \sim \text{Unif}(S) \) and \( S = \{(x, y) \in \mathbb{R}^{d+1} : y \leq \pi(x)\} \), then \( X \sim \pi \), i.e. the marginal distribution of \( X \) is \( \pi \).

**Proof.** \( (X, Y) \sim \text{Unif}(S) \)

\[ \Rightarrow \mathbb{P}(X \in (x, x + dx)) = \frac{\pi(x)dx}{\int \pi(v)dv} = \pi(x)dx \]

\( \Rightarrow \) the marginal density of \( X \) is \( \pi \). \hfill \square

Slice sampler to simulate from \( \pi(x) \): Use Gibbs sampling for Unif \((S)\).

Given \( x^{(t)} \),

1. Draw \( y^{(t+1)} \sim \text{Unif}[0, \pi(x^{(t)})] \);
2. Draw \( x^{(t+1)} \) uniformly from region \( A^{(t+1)} = \{x : \pi(x) \geq y^{(t+1)}\} \).

Then, \( (x^{(t)}, y^{(t)}) \sim \text{Unif}(S) \) when \( t \) is large, and by Lemma 1, \( x^{(t)} \sim \pi \), achieving the goal of sampling from \( \pi \).

If only know \( \pi(x) \propto f(x) \), replace \( \pi \) by \( f \) in the above algorithm.

**Example 8** (\( t_d \)-distribution). Use slice sampler to simulate from \( t \)-distribution with \( d \) degree of freedom:

\[ \pi(x) \propto (1 + x^2/d)^{-(d+1)/2} := f(x), \quad x \in \mathbb{R}. \]

Suppose the sample at iteration \( t \) is \( x_t \). The two steps to generate \( x_{t+1} \) are:
1. Draw \( y \sim \text{Unif}[0, f(x_t)] \), where \( f(x_t) = (1 + x_t^2/d)^{-(d+1)/2} \).

2. Draw \( x_{t+1} \) uniformly from the interval
\[
X_{t+1} = \{ x : f(x) \geq y \} = [-b(y), b(y)],
\]
where \( b(y) = \sqrt{d(y^{-2/(d+1)} - 1)} \). Note that \( \pm b(y) \) are the two roots of the quadratic equation \( f(x) = y \).

### 6.2. Metropolized Gibbs sampler

State space is discrete. \( x = (x_1, x_2, \cdots, x_d) \), \( x_i \) takes \( m_i \) possible values.

Random-scan Gibbs sampler, coordinate \( i \) is chosen.

Draw \( y_i (\neq x_i) \) with probability \( \frac{\pi(y_i | x_{[-i]})}{1 - \pi(x_i | x_{[-i]})} \).

Replace \( x_i \) by \( y_i \) with the M-H ratio: \( \min \left[ 1, \frac{1 - \pi(x_i | x_{[-i]})}{1 - \pi(y_i | x_{[-i]})} \right] \), because:
\[
\frac{\pi(y_i | x_{[-i]})}{\pi(x_i | x_{[-i]})} \cdot \frac{q(x_i, y_i)}{q(y_i, x_i)} = \frac{\pi(y_i | x_{[-i]})}{\pi(x_i | x_{[-i]})} \cdot \frac{1 - \pi(x_i | x_{[-i]})}{1 - \pi(y_i | x_{[-i]})} \cdot \frac{1 - \pi(x_i | x_{[-i]})}{1 - \pi(y_i | x_{[-i]})}.
\]

More efficient than random-scan Gibbs sampler.

### 6.3. Blocked Gibbs sampler

Partition \( \{1, \ldots, d\} \) into two blocks, \( A \) and \( B \): \( A \cup B = \{1, \ldots, d\} \) and \( A \cap B = \emptyset \).

Iterative sampling from \( [x_A | x_B] \) and \( [x_B | x_A] \).

**Example 9.** Ising model on a graph \( G = (V, E) \). \( V = \{1, \ldots, d\} \) is the vertex set. \( E \subset V \times V \) is the edge set: There is an edge between two vertices \( i, j \) iff \( (i, j) \in E \). Given \( G \), define a Boltzmann distribution for \((X_1, \ldots, X_d)\) at temperature \( T > 0 \):
\[
\pi(x_1, \ldots, x_d) \propto \exp \left( \frac{1}{T} \sum_{(i,j) \in E} x_i x_j \right), \quad x_i \in \{1,-1\}.
\]

This joint distribution implies the following conditional independence statements among \( X_1, \ldots, X_d \):
Theorem 3. Let $N_i = \{j : (i, j) \in E\}$ be the set of neighbors of vertex $i$ in the graph $G$. If $k \notin N_i$ and $k \neq i$, then

$$X_i \perp X_k \mid \{X_j : j \in N_i\}.$$

Proof. It follows from (6) that the conditional density of $X_i$ given $X_{[-i]}$ is

$$\pi(x_i \mid x_{[-i]}) \propto \exp \left( \frac{x_i}{T} \sum_{j \in N_i} x_j \right),$$

which only depends on $x_j, j \in N_i$.

1-d Ising model, $E = \{(1, 2), (2, 3), \ldots, (d - 1, d)\}$.

Complete graph, $E = \{(i, j) : i < j\}$.

Star topology, $E = \{(1, i) : i = 2, \ldots, d\}$. $X_1$ is the hub node (vertex) and $X_i \perp X_j \mid X_1$ for all $i \neq j \in \{2, \ldots, d\}$.

In this case, a two-block Gibbs sampler iterates between

1. Sample from $[x_1 \mid x_{[-1]}]$: Since

$$\pi(x_1 \mid x_2, \ldots, x_d) \propto \exp \left( \frac{1}{T} (x_2 + \ldots + x_d) x_1 \right),$$

for $x_1 \in \{1, -1\}$, after normalization we have

$$\pi(x_1 \mid x_2, \ldots, x_d) = \frac{\exp \left[ \frac{1}{T} (x_2 + \ldots + x_d) x_1 \right]}{\exp \left[ \frac{1}{T} (x_2 + \ldots + x_d) \right] + \exp \left[ \frac{-1}{T} (x_2 + \ldots + x_d) \right]},$$

$x_1 \in \{1, -1\}$.

2. Sample from $[x_{[-1]} \mid x_1]$, since $X_2, \ldots, X_d$ are independent given $X_1 = x_1$:

$$\pi(x_2, \ldots, x_d \mid x_1) \propto \prod_{j=2}^{d} \exp \left( \frac{x_1 x_j}{T} \right)$$

$$\propto \prod_{j=2}^{d} \pi(x_j \mid x_1), \quad x_j \in \{1, -1\}.$$

Draw $x_j$ from $[x_j \mid x_1]$ for each $j = 2, \ldots, d$ independently according to:

$$\pi(x_j \mid x_1) = \frac{\exp \left( \frac{1}{T} x_1 x_j \right)}{\exp \left( \frac{2}{T} \right) + \exp \left( \frac{-2}{T} \right)}, \quad x_j \in \{1, -1\}.$$
7. Bayesian Missing Data Problems

7.1. Introduction to Bayesian inference

Data: \( y_1, y_2, \ldots, y_n \) iid \( \sim f(y|\theta) \). Want to estimate \( \theta \).

\( y = (y_1, y_2, \ldots, y_n) \): observed data, \( \theta \): unknown parameter.

- Review of maximum likelihood estimate (MLE).
  Likelihood: \( L(\theta|y) \triangleq p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) \).
  MLE: \( (\hat{\theta})_{MLE} = \arg\max_{\theta} L(\theta|y) \).

- Bayesian estimate:
  1. Regard \( \theta \) as a random variable;
  2. Specify a prior distribution for \( \theta \), \( p(\theta) \);
  3. Find the posterior distribution for \( \theta \): \( p(\theta|y) = p(\theta|y_1, \ldots, y_n) \).

\[
p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} = \frac{p(\theta) \cdot \prod_{i=1}^{n} f(y_i|\theta)}{p(y)},
\]

where \( p(y) = \int p(\theta, y) d\theta \) is normalizing constant.

  More convenient to work with unnormalized posterior density:

\[
p(\theta|y) \propto p(\theta)p(y|\theta) = p(\theta) \cdot \prod_{i=1}^{n} f(y_i|\theta).
\]

- Construct Bayesian estimate:
  \( (\hat{\theta})_B : = \int \theta \cdot p(\theta|y) d\theta = \mathbb{E}(\theta|y) \): mean of posterior distribution.

  5. Monte Carlo simulation is usually used to draw from \( p(\theta | y) \).


Coin toss, \( \theta = \) probability of heads \( \in [0, 1] \).

Toss \( n \) times, observe \( X \) heads. Estimate \( \theta = ? \)

\( X | \theta \sim \text{Bin}(n, \theta) \). Likelihood \( P(X = x|\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \).

MLE: \( (\hat{\theta})_{MLE} = \frac{x}{n} \).

Bayesian inference, need to choose a prior distribution for \( \theta \):

(a) Flat prior (no information on \( \theta \)) \( \Rightarrow \theta \sim \text{Unif}(0, 1) \), i.e. \( p(\theta) = 1, \theta \in [0, 1] \).
Then posterior distribution:

\[ p(\theta|x) \propto p(\theta) \cdot p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad (\theta \text{ is the r.v.}) \]

\[ \Rightarrow [\theta|x] = \text{Beta}(x + 1, n - x + 1), \quad (\hat{\theta})_B = E(\theta|x) = \frac{x+1}{n+2}. \]

Beta(\alpha, \beta), pdf:

\[ \Gamma(\alpha+\beta) \Gamma(\alpha) \Gamma(\beta) \theta^{\alpha-1} (1-\theta)^{\beta-1} \]

E(\theta) = \frac{\alpha}{\alpha+\beta}

mode(\theta) = \frac{\alpha-1}{\alpha+\beta-2}

Shape of posterior distribution:

(b) Informative prior to incorporate prior knowledge.

Choose \( \theta \sim \text{Beta}(\alpha, \beta) \), \( p(\theta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1} \).

(For example, choose \( \alpha = \beta = 5 \), fair coin in prior.)

\[ \Rightarrow p(\theta|x) \propto p(\theta)p(x|\theta) \propto \theta^{x+\alpha-1} (1-\theta)^{n-x+\beta-1}. \]

\[ \Rightarrow [\theta|x] = \text{Beta}(x + \alpha, n - x + \beta). \quad (\alpha, \beta): \text{pseudo counts}. \]

\[ \hat{\theta}_B = \frac{x+\alpha}{n+\alpha+\beta} \]

**Example 11.** Multinomial distribution.

\( \theta = (\theta_1, \theta_2, \cdots, \theta_k) \), cell probabilities: \( \theta_j \geq 0, \sum_{j=1}^{k} \theta_j = 1. \)

Observations: \( x = (x_1, x_2, \cdots, x_k) \sim \text{M}(n, \theta), \sum x_j = n. \)

Likelihood \( p(x|\theta) \propto \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_k^{x_k}. \)

MLE \( (\hat{\theta}_j)_{MLE} = \frac{x_j}{n} \) for \( j = 1, \cdots, k. \)
Bayesian:

(a) Flat prior / non-informative.

\( \theta = (\theta_1, \cdots, \theta_k) \sim \text{Uniform distribution, i.e., } p(\theta) \propto 1. \)

Posterior: 

\[
p(\theta | x) \propto p(\theta) p(x | \theta) = \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_k^{x_k}. \quad (\sum \theta_j = 1.)
\]

\[\theta | x \sim \text{Dir}(x_1 + 1, x_2 + 1, \cdots, x_k + 1) \Rightarrow (\hat{\theta}_j)_B = \frac{x_j + 1}{n + k}, \quad j = 1, \cdots, k.\]

Dirichlet distribution: 

\( \theta = (\theta_1, \cdots, \theta_k), \theta_j \geq 0, \sum \theta_j = 1. \)

Then \( \theta \sim \text{Dir}(\alpha), \alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k). \) \( (\alpha_j > 0) \)

\[
p(\theta) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_k)} \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \cdots \theta_k^{\alpha_k-1}.
\]

\[E(\theta_j) = \frac{\alpha_j}{\alpha_1 + \alpha_2 + \cdots + \alpha_k}.\]

How to sample \( \theta \) from \( \text{Dir}(\alpha)? \)

1. Generate \( v_j \sim \text{Gamma}(\alpha_j, 1), \quad (pdf = \frac{1}{\Gamma(\alpha_j)} x^{\alpha_j-1} e^{-x}), \) for \( j = 1, 2, \cdots, k; \)
2. Define \( \hat{\theta}_j = \frac{v_j}{\sum_{i=1}^{k} v_i} = \frac{v_j}{v_1 + v_2 + \cdots + v_k}, \text{ for } j = 1, \cdots, k, \) then \( \theta = (\theta_1, \theta_2, \cdots, \theta_k) \sim \text{Dir}(\alpha_1, \alpha_2, \cdots, \alpha_k). \)

(b) Informative prior.

\( \theta \sim \text{Dir}(\alpha_1, \cdots, \alpha_k). \)

\[
\Rightarrow p(\theta | x) \propto \theta_1^{x_1+\alpha_1-1} \theta_2^{x_2+\alpha_2-1} \cdots \theta_k^{x_k+\alpha_k-1}.
\]

\[\theta | x \sim \text{Dir}(x_1 + \alpha_1, x_2 + \alpha_2, \cdots, x_k + \alpha_k), \alpha_0 = \sum_{j=1}^{k} \alpha_j.
\]

\( (\hat{\theta}_j)_B = \frac{x_j + \alpha_j}{n + \alpha_0}, \quad (\alpha_1, \cdots, \alpha_k): \text{pseudo counts}. \)

**Example 12.** Normal data with known variance.

\( y_1, \cdots, y_n \sim \text{iid } N(\theta, \sigma^2), \sigma^2 \text{ known.} \)

Likelihood

\[
p(y_1, \cdots, y_n | \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - \theta)^2 \right\}
\]

\[\propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2 \right\}.\]

MLE: \( (\hat{\theta})_{MLE} = \bar{y} = \frac{1}{n} \sum_i y_i. \)
Bayesian: flat prior \( p(\theta) \propto 1, \ \theta \in (-\infty, \infty) \).

Posterior distribution (\( \theta \) is the r.v.!) 

\[
p(\theta|y_1, \cdots, y_n) \propto p(\theta)p(y|\theta) \propto \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\theta - y_i)^2 \right\}.
\]

\[
\Rightarrow \sum_{i=1}^{n} (\theta - y_i)^2 = n(\theta - \bar{y})^2 + \sum_{i=1}^{n} y_i^2 - n\bar{y}^2
\]

\[
p(\theta|y_1, \cdots, y_n) \propto \exp \left\{ -\frac{1}{2\sigma^2} n(\theta - \bar{y})^2 \right\} = \exp \left\{ -\frac{(\theta - \bar{y})^2}{2\sigma^2/n} \right\}
\]

\[
\Rightarrow [\theta|y_1, \cdots, y_n] = N(\bar{y}, \sigma^2/n) \text{ and } (\hat{\theta})_B = E(\theta|y_1, \cdots, y_n) = \bar{y}.
\]

7.2. Data augmentation

\[y_1, y_2, \cdots, y_n \overset{iid}{\sim} f(y|\theta), \ y_i = (y_{i1}, y_{i2}, \cdots, y_{ip}) \in \mathbb{R}^p.\]

Data matrix \( Y = (y_{ij})_{n \times p} \) contains missing elements.

\[
Y = \begin{bmatrix} Y_{\text{obs}} & Y_{\text{mis}} \end{bmatrix}_{n \times p}
\]

- ?: missing. (e.g. \( y_{12}, y_{np}, \cdots \))
- \( \checkmark \): observed. (e.g. \( y_{11}, y_{1p}, \cdots \))
- \( Y_{\text{obs}} \): observed elements of \( Y \) (observed data).
- \( Y_{\text{mis}} \): missing elements of \( Y \) (missing data).
- \( Y = (Y_{\text{obs}}, Y_{\text{mis}}) \): complete data.

Bayesian inference: (1) estimate \( \theta \); (2) predict missing data.
Calculate the joint posterior distribution of \((\theta, Y_{\text{mis}})\):

\[
p(\theta, Y_{\text{mis}} | Y_{\text{obs}}) \propto p(\theta) p(Y_{\text{obs}}, Y_{\text{mis}} | \theta),
\]

where \(p(\theta)\) is the prior for \(\theta\) and \(p(Y_{\text{obs}}, Y_{\text{mis}} | \theta)\) is complete-data likelihood.

Usually there are no closed-form formulas for posterior mean or quantiles of

\[
p(\theta | Y_{\text{obs}}) \propto \int p(Y_{\text{obs}}, Y_{\text{mis}} | \theta) dY_{\text{mis}}
\]

which is a mixture of complete-data posterior \(p(\theta | Y)\). We need to draw \((\theta, Y_{\text{mis}}) | Y_{\text{obs}}\).

Two block Gibbs sample for \(p(\theta, Y_{\text{mis}} | Y_{\text{obs}})\):

1. Given \(\theta^{(t)}\), sample \(Y_{\text{mis}}^{(t+1)} \sim p(Y_{\text{mis}} | Y_{\text{obs}}, \theta^{(t)})\);

2. Given \(Y_{\text{mis}}^{(t+1)}\), sample \(\theta^{(t+1)} \sim p(\theta | Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)}) = p(\theta | Y^{(t+1)})\) (complete-data posterior).

\[
\Rightarrow \{ (\theta^{(t)}, Y_{\text{mis}}^{(t)}) : t = 1, 2, \ldots \}: \text{Markov chain with stationary distribution} \ p(\theta, Y_{\text{mis}} | Y_{\text{obs}}).
\]

**Example 13.** \(x_1, x_2, \ldots, x_n \overset{\text{iid}}{\sim} \text{Discrete } (\theta_1, \theta_2, \theta_3)\):

\[
\mathbb{P}(x_i = k) = \theta_k, \ k = 1, 2, 3.
\]
Prior: $\theta \sim \text{Dir}(\alpha_1, \alpha_2, \alpha_3), \quad (\theta_1 + \theta_2 + \theta_3 = 1)$

$$p(\theta_1, \theta_2, \theta_3) \propto \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \theta_3^{\alpha_3-1}.$$ 

Missing data in $x_1, x_2, x_3$, and $Y_{\text{obs}} = (x_1 \neq 1, x_2 \neq 2, x_3 \neq 3, x_4, \ldots, x_n)$.

$$p(\theta, x_1, x_2, x_3 | x_4, \ldots, x_n) \propto p(\theta) p(x_1, x_2, x_3, x_4, \ldots, x_n | \theta)$$

$$\propto \left(3 \prod_{j=1}^{3} \theta_j^{\alpha_j-1}\right) \left(3 \prod_{i=1}^{3} p(x_i | \theta)\right) \left(3 \prod_{j=1}^{3} C_j^{(\text{obs})}\right)$$

$$\propto \left(3 \prod_{j=1}^{3} \theta_j^{\alpha_j + C_j^{(\text{obs})} - 1}\right) \left(3 \prod_{i=1}^{3} p(x_i | \theta)\right) ,$$

where $C_j^{\text{obs}} = \sum_{i=4}^{n} I(x_i = j)$: observed counts for the $j$th category from $x_4$ to $x_n$.

1. Given $\theta = (\theta_1, \theta_2, \theta_3)$, $p(x_1 = j | \theta) \propto \theta_j$ for $j = 1, 2, 3$,

$$\Rightarrow p(x_1 = j | x_1 \neq 1, \theta) = \frac{\theta_j}{\theta_2 + \theta_3}, \quad j = 2, 3.$$ 

Similarly,

$$p(x_2 = j | x_2 \neq 2, \theta) = \frac{\theta_j}{\theta_1 + \theta_3}, \quad j = 1, 3.$$ 

$$p(x_3 = j | x_3 \neq 3, \theta) = \frac{\theta_j}{\theta_1 + \theta_2}, \quad j = 1, 2.$$ 

Draw $x_1, x_2, x_3$ independently according to the above conditional probabilities.

2. Given $(x_1, x_2, x_3)$, $C_j^{(\text{mis})} = \sum_{i=1}^{3} I(x_i = j)$,

then $p(\theta | x_1, \ldots, x_n) \propto \prod_{j=1}^{3} \theta_j^{C_j^{(\text{obs})} + C_j^{(\text{mis})} + \alpha_j - 1}$. Draw $\theta$ from

$$\theta | x \sim \text{Dir}(C_1^{(\text{obs})} + C_1^{(\text{mis})} + \alpha_1, C_2^{(\text{obs})} + C_2^{(\text{mis})} + \alpha_2, C_3^{(\text{obs})} + C_3^{(\text{mis})} + \alpha_3),$$

where $x = (x_1, \ldots, x_n)$ is complete data.

Iterate between steps 1 and 2 to generate $(\theta^{(t)}, x_1^{(t)}, x_2^{(t)}, x_3^{(t)})$ for $t = 1, \ldots, m$.

Bayesian estimates: $\hat{\theta}_B \approx \frac{1}{m} \sum_t \theta^{(t)}$ and histogram of $\theta^{(t)}$. 
Example 14. $y_1, y_2, \ldots, y_n \overset{iid}{\sim} \mathcal{N}_2(\mu, \Sigma), y_i = (y_{i1}, y_{i2})$.

\[
\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma_{\text{known}} = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\
-\rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}.
\]

Improper flat prior: $p(\mu) \propto 1$.

Missing data $Y_{\text{mis}} = (y_{11}, y_{22})$ and observed data $Y_{\text{obs}} = (y_{12}, y_{21}, y_3, \ldots, y_n)$.

Data augmentation for this problem:

1. Given $\mu$, sample $y_{11}$ and $y_{22}$, $[y_{11}|y_{12}, \mu, \Sigma] \sim$? Recall $y_1 = (y_{11}, y_{12})$.

\[
p(y_{11}|y_{12}, \mu, \Sigma) \propto p(y_{11}, y_{12}|\mu, \Sigma) \propto \exp\left\{-\frac{1}{2}(y_1 - \mu)^T \Sigma^{-1} (y_1 - \mu)\right\}
\]

\[
= \exp\left\{-\frac{1}{2(1-\rho^2)} \left[ \frac{(y_{11} - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(y_{11} - \mu_1)(y_{12} - \mu_2)}{\sigma_1 \sigma_2} + \frac{(y_{12} - \mu_2)^2}{\sigma_2^2} \right] \right\}
\]

\[
\propto \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_1^2} \left[ (y_{11} - \mu_1)^2 - \frac{2\rho \sigma_1}{\sigma_2^2} (y_{12} - \mu_2)(y_{11} - \mu_1) \right] \right\}
\]

\[
= \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_1^2} \left[ y_{11} - \mu_1 - \frac{\rho \sigma_1}{\sigma_2} (y_{12} - \mu_2) \right]^2 + C \right\}.
\]

\[\therefore y_{11}|y_{12}, \mu, \Sigma \sim \mathcal{N}\left( \mu_1 + \frac{\rho \sigma_1}{\sigma_2} (y_{12} - \mu_2), (1-\rho^2)\sigma_1^2 \right).\]

Similarly, $y_{22}|y_{21}, \mu, \Sigma \sim \mathcal{N}\left( \mu_2 + \frac{\rho \sigma_1}{\sigma_2} (y_{21} - \mu_1), (1-\rho^2)\sigma_2^2 \right)$.

Given $\mu$, draw $y_{11}$ and $y_{22}$ independently from the two normal distributions.
2. Given $y_{11}$ and $y_{22}$, sample $\mu$?

$$p(\mu | y_1, y_2, \cdots, y_n, \Sigma) \propto p(y_1, \cdots, y_n | \mu, \Sigma) = (2\pi | \Sigma |)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \right]$$

$$\propto \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^T \Sigma^{-1} (y_i - \mu) \right].$$

Let $\bar{y} = \sum_i y_i / n$.

$$\sum_i (\mu - y_i)^T \Sigma^{-1} (\mu - y_i)$$

$$= \sum_i (\mu - \bar{y} + \bar{y} - y_i)^T \Sigma^{-1} (\mu - \bar{y} + \bar{y} - y_i)$$

$$= \sum_i \left[ (\mu - \bar{y})^T \Sigma^{-1} (\mu - \bar{y}) + 2(\mu - \bar{y})^T \Sigma^{-1} (\bar{y} - y_i) + (\bar{y} - y_i)^T \Sigma^{-1} (\bar{y} - y_i) \right]$$

$$= n(\mu - \bar{y})^T \Sigma^{-1} (\mu - \bar{y}) + C.$$ 

Therefore, $\mu | y_1, \cdots, y_n \sim N_2(\bar{y}, \frac{1}{n} \Sigma)$.

Iterate between steps 1 and 2.