Gibbs Sampling and Markov Chains

王璐

对 Bayesian 模型，一般很难直接从参数的联合后验分布 (joint posterior distribution) 抽样，但有时从每个参数的 full conditional distribution 中抽样较容易。这种情况下，可以使用 Gibbs sampler 得到很多近似服从后验分布的样本，然后用这些样本近似描述后验分布。Gibbs sampler 是一种迭代抽样算法，随着样本数增加，样本的分布会收敛到目标分布，但是它产生的样本之间有相关性。我们首先以一个简单的 Bayesian 模型估计为例介绍如何使用 Gibbs sampler.

1 A Bayesian Normal Model

使用 Bayesian 方法分析数据一般有以下三要素 (Hoff, 2009):

1. 模型设定：为数据的抽样分布设定具体形式 $p(\boldsymbol{y} \mid \boldsymbol{\theta})$, 通常需要引入一些参数 $\boldsymbol{\theta}$。比如，假设数据独立地服从正态分布：

$$Y_i \overset{iid}{\sim} N(\mu, \phi^{-1}), \quad i = 1, \ldots, n. \tag{1}$$

   这里我们用参数 ϕ^{-1} 表示正态分布的方差，ϕ 被称为 precision.

2. 设定参数的先验分布 (prior): 参数 $\boldsymbol{\theta}$ 的先验分布 $p(\boldsymbol{\theta})$ 一般是主观设定的，可以加入参数的 prior information, 其 support 应涵盖参数所有可能的取值范围。例如，可以为模型(1)的参数设定如下的先验分布:

$$\begin{aligned} \mu &\sim N(\mu_0, \tau_0^2) \\ \phi &\sim \text{Gam}(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned} \tag{2}$$

   其中 μ_0, τ_0^2, ν_0, σ_0^2 都是确定的常数。

3. 计算参数的后验分布并做统计推断：得到参数 $\boldsymbol{\theta}$ 的后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 后，可以估计参数的后验期望 $E(\boldsymbol{\theta} \mid \boldsymbol{y})$、后验方差 $Var(\boldsymbol{\theta} \mid \boldsymbol{y})$、置信区间等。参数的后验分布可如下计算：

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta})d\boldsymbol{\theta}} \propto p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta}) \tag{3}$$

   但 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 对应的分布一般很难识别或很难直接对其抽样。

如果我们为模型(1)选取如下的 conjugate prior:

$$\phi \sim \text{Gam}(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$$
$$\mu \mid \phi \sim N(\mu_0, \frac{1}{\kappa_0\phi})$$

仿照 normal-inverse-Wishart 分布的计算，以及等式 $\sum_{i=1}^n (y_i - \bar{y})^2 = \left(\sum_{i=1}^n y_i^2\right) - n\bar{y}^2$，可得 (μ, ϕ) 的联合后验分布为

$$p(\mu, \phi \mid y_1, \ldots, y_n) = p(\mu \mid \phi, y_1, \ldots, y_n)p(\phi \mid y_1, \ldots, y_n)$$

其中 $p(\phi \mid y_1, \ldots, y_n)$ 是一个 gamma density, $p(\mu \mid \phi, y_1, \ldots, y_n)$ 是一个 normal density, 具体形式为:

$$\phi \mid y_1, \ldots, y_n \sim \text{Gam}(\frac{\nu_n}{2}, \frac{S_n}{2})$$
$$\mu \mid \phi, y_1, \ldots, y_n \sim N(\mu_n, \frac{1}{\kappa_n\phi})$$

其中

$$\nu_n = \nu_0 + n$$
$$S_n = \nu_0\sigma_0^2 + \frac{n\kappa_0}{\kappa_0 + n}(\mu_0 - \bar{y})^2 + \sum_{i=1}^n (y_i - \bar{y})^2$$
$$\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}$$
$$\kappa_n = \kappa_0 + n$$

如果想计算 μ 的边际后验期望 (marginal posterior mean) $E(\mu \mid y_1, \ldots, y_n)$，可以采用如下的 Monte Carlo 方法:

(1) 抽取 $\phi^{(s)} \sim \text{Gam}(\nu_n/2, S_n/2)$, $s = 1, \ldots, T$.

(2) 对每个 $\phi^{(s)}$, 抽取 $\mu^{(s)} \mid \phi^{(s)} \sim N(\mu_n, (\kappa_n\phi^{(s)})^{-1})$, $s = 1, \ldots, T$.

则 $E(\mu \mid y_1, \ldots, y_n) \approx \sum_{s=1}^T \mu^{(s)}/T$.

如果我们选取(2)中的 prior, 则 ϕ 的边际后验分布既不是 Gamma 分布，也不是任何常见的容易抽样的分布 (练习)。这种情况下，可以尝试用数值方法得到 (μ, ϕ) 近似的联合后验分布:

- 首先为各参数选取足够大的取值范围 $\mu \in [\mu_L, \mu_H]$, $\phi \in [\phi_L, \phi_H] \subseteq (0, \infty)$.
- 然后对区域 $[\mu_L, \mu_H] \times [\phi_L, \phi_H]$ 做网格离散，比如在区间 $[\mu_L, \mu_H]$ 和 $[\phi_L, \phi_H]$ 上各取等距的 1000 个点 $\{\mu_1, \ldots, \mu_{1000}\}$, $\{\phi_1, \ldots, \phi_{1000}\}$.

- 根据(3)，点 (μ_i, ϕ_j) 处的后验概率密度为

$$p(\mu_i, \phi_j \mid y_1, \ldots, y_n) \propto p(\mu_i, \phi_j) p(y_1, \ldots, y_n \mid \mu_i, \phi_j),$$

因此网格中每个点近似的后验概率为：

$$P(\mu_i, \phi_j \mid y_1, \ldots, y_n) = \frac{p(\mu_i, \phi_j) p(y_1, \ldots, y_n \mid \mu_i, \phi_j)}{\sum_{i=1}^{1000} \sum_{j=1}^{1000} p(\mu_i, \phi_j) p(y_1, \ldots, y_n \mid \mu_i, \phi_j)}. \tag{4}$$

由(4)定义的离散分布可以近似 (μ, ϕ) 的联合后验分布。从该离散分布，我们也可以估计出各参数的边际后验期望和置信区间等。但是网格中大部分点的后验概率都很接近 0，造成计算的浪费。而且该方法只适用参数较少的情况，随着参数维度 p 的增加，网格点的数目呈指数增长，因此在高维参数情形下该方法不可行。这促使人们发明了更高效地获取参数后验分布样本的 Gibbs sampler 方法。

Gibbs sampler 需要计算每个参数的 full conditional distribution.

- 模型(1)-(2) 中 μ 的 full conditional density:

$$\begin{aligned} p(\mu \mid \phi, y_1, \ldots, y_n) &\propto p(\mu) p(\phi) p(y_1, \ldots, y_n \mid \mu, \phi) \\ &\propto \exp\left\{ -\frac{1}{2\tau_0^2} (\mu - \mu_0)^2 \right\} \exp\left\{ -\frac{\phi}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left[(\tau_0^{-2} + n\phi)\mu^2 - 2\mu(\tau_0^{-2}\mu_0 + \phi \sum_{i=1}^{n} y_i) \right] \right\} \end{aligned}$$

因此 $\mu \mid \phi, y_1, \ldots, y_n \sim N(\mu_n, \tau_n^2)$, 其中 $\mu_n = (\tau_0^{-2}\mu_0 + \phi \sum_{i=1}^{n} y_i)/(\tau_0^{-2} + n\phi)$, $\tau_n^2 = (\tau_0^{-2} + n\phi)^{-1}$.

- 模型(1)-(2) 中 ϕ 的 full conditional density:

$$\begin{aligned} p(\phi \mid \mu, y_1, \ldots, y_n) &\propto p(\mu) p(\phi) p(y_1, \ldots, y_n \mid \mu, \phi) \\ &\propto \phi^{\nu_0/2 - 1} \exp\left(-\frac{\nu_0 \sigma_0^2}{2} \phi \right) \phi^{n/2} \exp\left\{ -\frac{\phi}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \\ &\propto \phi^{(\nu_0 + n)/2 - 1} \exp\left\{ -\phi \left(\nu_0 \sigma_0^2 + \sum_{i=1}^{n} (y_i - \mu)^2 \right) / 2 \right\} \end{aligned}$$

因此 $\phi \mid \mu, y_1, \ldots, y_n \sim \mathrm{Gam}(\nu_n/2, S_n/2)$, 其中 $\nu_n = \nu_0 + n$, $S_n = \nu_0 \sigma_0^2 + \sum_{i=1}^{n} (y_i - \mu)^2$.

那么如何利用 μ 和 ϕ 的 full conditional distributions 得到 (μ, ϕ) 的联合后验分布的样本? 假设 $\phi^{(1)}$ 是边际后验分布 $p(\phi \mid y_1, \ldots, y_n)$ 的一个样本。给定 $\phi^{(1)}$, 从 μ 的 full conditional distribution 抽样：

$$\mu^{(1)} \sim p(\mu \mid \phi^{(1)}, y_1, \ldots, y_n)$$

则 $(\mu^{(1)}, \phi^{(1)})$ 可以看作联合后验分布 $p(\mu, \phi \mid y_1, \ldots, y_n)$ 的一个样本, 且 $\mu^{(1)}$ 可以看作边际分布 $p(\mu \mid y_1, \ldots, y_n)$ 的一个样本。因此给定 $\mu^{(1)}$, 从 ϕ 的 full conditional distribution 抽取

$$\phi^{(2)} \sim p(\phi \mid \mu^{(1)}, y_1, \ldots, y_n)$$

则 $(\mu^{(1)}, \phi^{(2)})$ 又可以看作联合后验分布 $p(\mu, \phi \mid y_1, \ldots, y_n)$ 的一个样本, $\phi^{(2)}$ 可看作边际分布 $p(\phi \mid y_1, \ldots, y_n)$ 的一个样本, 继续用来产生 $\mu^{(2)}$, 以此类推。

因此只要给定 μ 或 ϕ 的一个初始值, 然后轮流从 μ 和 ϕ 的 full conditional distributions 抽样, 就可以得到一列来自联合后验分布 $p(\mu, \phi \mid y_1, \ldots, y_n)$ 的样本 $\{(\mu^{(s)}, \phi^{(s)}) : s = 1, \ldots, T\}$, 且 $\{\mu^{(s)}\}_{s=1}^T$ 和 $\{\phi^{(s)}\}_{s=1}^T$ 可看作分别来自 μ 和 ϕ 的边际后验分布的样本。这种方法被称为 **Gibbs sampler**. 模型(1)-(2)对应的 Gibbs sampler 的 R code 如下:

```
## create data
set.seed(1)
n = 100
y = rnorm(n, mean=-5, sd=2)

# specify prior parameters
mu0 = 0
tau02 = 10
nu0 = 4
sigma02 = 10

## posterior samples ----------
ns = 1000 # number of samples
mu_samples = numeric(ns)
phi_samples = numeric(ns)
# starting values
mu = mean(y)
phi = 1/var(y)
mu_samples[1] = mu
phi_samples[1] = phi
# intermediate variables
sum_y = sum(y)

# Gibbs sampling
set.seed(10)
```



```
for (s in 2:ns){
  # generate mu
  mu_n = (mu0/tau02 + phi*sum_y)/(1/tau02 + n*phi)
  tau2_n = 1/(1/tau02 + n*phi)
  mu = rnorm(1, mean = mu_n, sd = sqrt(tau2_n))
  mu_samples[s] = mu

  # generate phi
  nu_n = nu0 + n
  Sn = nu0*sigma02 + sum((y-mu)^2)
  phi = rgamma(1, shape = nu_n/2, rate = Sn/2)
  phi_samples[s] = phi
}
```

图1展示了上述得到的 μ 和 ϕ 的后验样本随迭代步数的移动，称为 traceplot. 从该图可以看到，本例中 Gibbs sampler 得到的样本在参数空间中 “移动” 的很快，此时称样本的 mixing 很好，或 mixing 速度很快。这表明样本之间的相关性较小，此时样本均值可以很好地近似后验分布的期望。

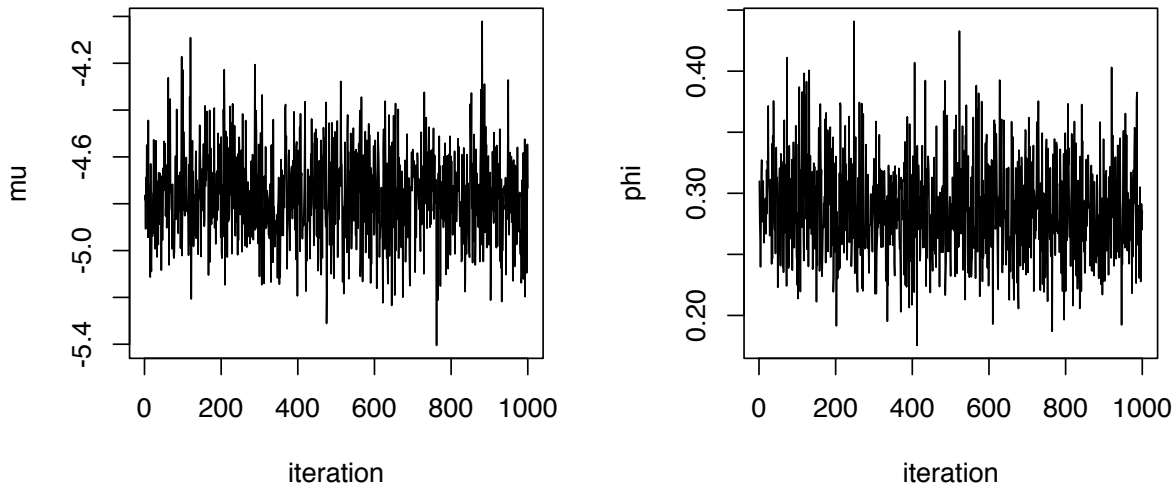

Figure 1: μ (左) 和 ϕ (右) 后验样本的 traceplots.

样本之间的相关性如何影响样本均值对目标期望的近似? 假设 $\{\theta^{(s)} : s = 1, \ldots, T\}$ 是由 Gibbs sampling 得到的一列服从目标分布 $p(\theta)$ 的样本，此时样本均值 $\bar{\theta} = \sum_{s=1}^{T} \theta^{(s)}/T$ 到目标期望 $E(\theta) = \int \theta p(\theta) d\theta$ 距离平方的期望，即 $\bar{\theta}$ 的方差为

$$
\begin{aligned}
Var_G(\bar{\theta}) &= E\left[\left(\bar{\theta} - E(\theta)\right)^2\right] \\
&= E\left[\left(\frac{1}{T}\sum_{s=1}^{T}\left(\theta^{(s)} - E(\theta)\right)\right)^2\right] \\
&= \frac{1}{T^2} E\left[\sum_{s=1}^{T}\left(\theta^{(s)} - E(\theta)\right)^2 + \sum_{s=1}^{T}\sum_{t\neq s}\left(\theta^{(s)} - E(\theta)\right)\left(\theta^{(t)} - E(\theta)\right)\right] \\
&= \frac{1}{T^2}\sum_{s=1}^{T} E\left[\left(\theta^{(s)} - E(\theta)\right)^2\right] + \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t\neq s} E\left[\left(\theta^{(s)} - E(\theta)\right)\left(\theta^{(t)} - E(\theta)\right)\right] \\
&= \frac{1}{T} Var(\theta) + \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t\neq s} E\left[\left(\theta^{(s)} - E(\theta)\right)\left(\theta^{(t)} - E(\theta)\right)\right] \\
&= Var_{MC}(\bar{\theta}) + \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t\neq s} E\left[\left(\theta^{(s)} - E(\theta)\right)\left(\theta^{(t)} - E(\theta)\right)\right]
\end{aligned}
$$

其中 $Var_{MC}(\bar{\theta})$ 是 $p(\theta)$ 独立同分布的样本的均值对应的方差。等式右边第二项取决于样本 $\{\theta^{(s)}\}_{s=1}^{T}$ 之间的相关性。由于 Gibbs sampler 产生的样本之间的相关性一般为正，这一项通常大于 0，因此 $Var_G(\bar{\theta}) > Var_{MC}(\bar{\theta})$, 即 Gibbs sampling 的样本均值到目标期望的距离平均会大于独立的 Monte Carlo 样本的均值到目标期望的距离。且 Gibbs sampling 样本之间的相关性越高，$Var_G(\bar{\theta})$ 就越大，均值的近似效果越差。一个衡量 Gibbs sampling 样本相关性的指标是 *effective sample size* (ESS), 定义为

$$ESS = \frac{Var(\theta)}{Var_G(\bar{\theta})}$$

ESS 可以理解为：为达到与 Gibbs sampling 样本估计量相同精度所需的独立样本的个数。R package `mcmcse` 的函数 `ess()` 可以给出 Gibbs sampling 样本的 ESS.

```
# diagnostics
library(mcmcse)
plot(mu_samples, type='l', xlab='iteration', ylab = 'mu')
plot(phi_samples, type='l', xlab='iteration', ylab = 'phi')
acf(mu_samples) # autocorrelation
acf(phi_samples)
ess(mu_samples)
ess(phi_samples)
```


对上述 Bayesian 模型(1)-(2)，Gibbs sampler 得到的 μ 的 1000 个后验样本的 ESS 为 930，ϕ 的后验样本的 ESS 为 961，说明 Gibbs sampler 在该模型中表现得非常好。利用这些后验样本，我们可以估计参数的后验期望和后验置信区间 (credible interval)：μ 的后验均值为 $\bar{\mu} = -4.77$, 95% credible interval 为 (-5.14, -4.38); ϕ 的后验均值为 $\bar{\phi} = 0.29$, 95% credible interval 为 (0.22, 0.37).

2 Gibbs Sampling

本节我们给出一般的 Gibbs sampling 算法。假设模型的参数向量为 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, 其中每个分量 θ_j 可以是一个数或一个向量，$j = 1, \ldots, p$. 抽样的目标分布为 $p(\boldsymbol{\theta})$, 但很难直接从 $p(\boldsymbol{\theta})$ 中抽样，比如在上述 Bayesian 模型 (1)-(2)中，目标分布为 $p(\mu, \phi \mid y_1, \ldots, y_n)$. 如果我们知道 $\boldsymbol{\theta}$ 每个分量的 full conditional distribution $p(\theta_j \mid \{\theta_k\}_{k \neq j})$, $j = 1, \ldots, p$, 给定 $\boldsymbol{\theta}$ 的一个初始值 $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_p^{(0)})$, Gibbs sampler 可以如下从当前样本 $\boldsymbol{\theta}^{(s-1)}$ 产生新的样本 $\boldsymbol{\theta}^{(s)}$:

1. 抽取 $\theta_1^{(s)} \sim p(\theta_1 \mid \theta_2^{(s-1)}, \theta_3^{(s-1)}, \ldots, \theta_p^{(s-1)})$
2. 抽取 $\theta_2^{(s)} \sim p(\theta_2 \mid \theta_1^{(s)}, \theta_3^{(s-1)}, \ldots, \theta_p^{(s-1)})$
3. 抽取 $\theta_3^{(s)} \sim p(\theta_3 \mid \theta_1^{(s)}, \theta_2^{(s)}, \theta_4^{(s-1)}, \ldots, \theta_p^{(s-1)})$

   $\vdots$

p. 抽取 $\theta_p^{(s)} \sim p(\theta_p \mid \theta_1^{(s)}, \theta_2^{(s)}, \ldots, \theta_{p-1}^{(s)})$.

在上述过程中，如果能将 $\boldsymbol{\theta}$ 的一些一元分量 “合并” 成一个子向量进行更新，就可以减少样本之间的相关性，这种方法称为 *block sampling*.

不断重复上述过程，Gibbs sampler 可以产生一列不独立的样本 $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(T)}$, 且每个样本 $\boldsymbol{\theta}^{(s)}$ 与之前的样本 $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(s-1)}$ 的相关性只取决于 $\boldsymbol{\theta}^{(s-1)}$, 即给定 $\boldsymbol{\theta}^{(s-1)}$, $\boldsymbol{\theta}^{(s)}$ 与 $\boldsymbol{\theta}^{(0)}, \ldots, \boldsymbol{\theta}^{(s-2)}$ 条件独立，这被称为 Markov 性，因此称这一列样本是一条 **Markov chain**.

当满足一些条件时（后面会介绍），从任何初始值 $\boldsymbol{\theta}^{(0)}$ 出发，Gibbs sampler 产生的样本 $\boldsymbol{\theta}^{(s)}$ 的分布在 $s \to \infty$ 时收敛到目标分布 $p(\boldsymbol{\theta})$, 即

$$P(\boldsymbol{\theta}^{(s)} \in A) \to \int_A p(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

并且对任意可积函数 g 有

$$\frac{1}{T} \sum_{s=1}^{T} g(\boldsymbol{\theta}^{(s)}) \to E\left[g(\boldsymbol{\theta})\right] = \int g(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad s \to \infty.$$

这说明我们可以用 $\{g(\boldsymbol{\theta}^{(s)})\}$ 的样本均值近似 $E\left[g(\boldsymbol{\theta})\right]$, 这与 Monte Carlo 方法相似，因此上述过程被称为 **Markov chain Monte Carlo (MCMC)** 方法。

3 Markov Chains

本节介绍 Markov chains 的一些基本理论。

从一个 p 维状态空间 (state space) $\mathcal{X}$ 上按如下方式随机抽取一列向量：给定一个初始值 $\boldsymbol{x}^{(0)} \in \mathcal{X}$, 按照某个条件分布 $p(\boldsymbol{x} \mid \boldsymbol{x}')$ 依次抽取

$$\boldsymbol{x}^{(t)} \sim p(\boldsymbol{x} \mid \boldsymbol{x}^{(t-1)}) \quad \text{且} \quad \boldsymbol{x}^{(t)} \perp\!\!\!\perp \boldsymbol{x}^{(t-k)} \mid \boldsymbol{x}^{(t-1)},\ k > 1.$$

称这样得到的序列 $\{\boldsymbol{x}^{(t)}\}$ 为状态空间 $\mathcal{X}$ 上的一条 Markov chain, 或一阶 Markov process.

与 Markov chain 有关的一些基本概念：

- **Transition densities**. 在 Markov chain 中, 对任意正整数 k, 定义如下的 **k-step transition density**
  - $p^1(\boldsymbol{x}^{(t+1)} \mid \boldsymbol{x}^{(t)}) = p(\boldsymbol{x}^{(t+1)} \mid \boldsymbol{x}^{(t)})$
  - $p^2(\boldsymbol{x}^{(t+2)} \mid \boldsymbol{x}^{(t)}) = \int_{\mathcal{X}} p(\boldsymbol{x}^{(t+2)} \mid \boldsymbol{x}^{(t+1)}) p(\boldsymbol{x}^{(t+1)} \mid \boldsymbol{x}^{(t)}) d\boldsymbol{x}^{(t+1)}$
  - $p^3(\boldsymbol{x}^{(t+3)} \mid \boldsymbol{x}^{(t)}) = \int_{\mathcal{X}} p(\boldsymbol{x}^{(t+3)} \mid \boldsymbol{x}^{(t+2)}) p^2(\boldsymbol{x}^{(t+2)} \mid \boldsymbol{x}^{(t)}) d\boldsymbol{x}^{(t+2)}$
  - $\vdots$
  - $p^k(\boldsymbol{x}^{(t+k)} \mid \boldsymbol{x}^{(t)}) = \int_{\mathcal{X}} p(\boldsymbol{x}^{(t+k)} \mid \boldsymbol{x}^{(t+k-1)}) p^{k-1}(\boldsymbol{x}^{(t+k-1)} \mid \boldsymbol{x}^{(t)}) d\boldsymbol{x}^{(t+k-1)}$

  如果 transition density $p(\boldsymbol{x} \mid \boldsymbol{x}') > 0, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, 则对任意正整数 k, k-step transition density $p^k(\boldsymbol{x} \mid \boldsymbol{x}') > 0, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. 这说明从 $\boldsymbol{x}^{(t)}$ 出发，该过程在 k 步后到达任意 $\boldsymbol{x}^{(t+k)} \in \mathcal{X}$ 的概率密度都为正，即该过程可以 “漫游” 到状态空间的任何地方。通常具有该性质的 Markov process 会收敛到唯一的极限分布。

- **Stationary distribution**. 如果存在一个分布 $\pi(\boldsymbol{x})$ 使 Markov process 满足

$$\pi(\boldsymbol{x}) = \int_{\mathcal{X}} p(\boldsymbol{x} \mid \boldsymbol{x}') \pi(\boldsymbol{x}') d\boldsymbol{x}'. \tag{5}$$

  称分布 $\pi(\boldsymbol{x})$ 为该 Markov process 的一个 **stationary distribution**. (5)表明如果 Markov chain 的当前状态 $\boldsymbol{x}'$ 服从 stationary distribution $\pi(\cdot)$, 按照 transition density $p(\boldsymbol{x} \mid \boldsymbol{x}')$ 产生的新状态 $\boldsymbol{x}$ 的边际分布依然是 $\pi(\cdot)$.

  由(5)得

$$\pi(\boldsymbol{x}) = \int_{\mathcal{X}} p^k(\boldsymbol{x} \mid \boldsymbol{x}') \pi(\boldsymbol{x}') d\boldsymbol{x}', \quad \forall k \in \mathbb{N}^+. \tag{6}$$

  (6)表明，如果 Markov chain 的初始值服从 stationary distribution, 整个过程的边际分布会永远 “保持” stationary distribution.

在 Bayesian 分析中，Gibbs sampler 产生的 Markov chain 的 transition density 是

$$p(\boldsymbol{\theta}^{(s)} \mid \boldsymbol{\theta}^{(s-1)}) = p(\theta_1^{(s)} \mid \theta_2^{(s-1)}, \ldots, \theta_p^{(s-1)}) p(\theta_2^{(s)} \mid \theta_1^{(s)}, \theta_3^{(s-1)}, \ldots, \theta_p^{(s-1)}) \cdots p(\theta_p^{(s)} \mid \theta_1^{(s)}, \ldots, \theta_{p-1}^{(s)}) \tag{7}$$

其中每个分量的条件分布都是该参数的 full conditional distribution. 如果 $\forall \boldsymbol{\theta}^{(s-1)}, \boldsymbol{\theta}^{(s)}$, $p(\boldsymbol{\theta}^{(s)} \mid \boldsymbol{\theta}^{(s-1)}) > 0$, 该过程会收敛到唯一的 stationary distribution, 且这个分布是参数的后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. $\boldsymbol{\theta} \in \mathbb{R}^2$ 的证明如下，对任意 p 维向量 $\boldsymbol{\theta}$ 也可类似证明。

Proof. 假设 $\boldsymbol{\theta}^{(s-1)} = (\theta_1^{(s-1)}, \theta_2^{(s-1)})$ 服从后验分布 $p(\theta_1, \theta_2 \mid \boldsymbol{y})$, 则由 Gibbs sampler 产生的 $\boldsymbol{\theta}^{(s)} = (\theta_1^{(s)}, \theta_2^{(s)})$ 的边际分布为

$$\begin{aligned}
&\int\int p(\theta_1^{(s)}, \theta_2^{(s)} \mid \theta_1^{(s-1)}, \theta_2^{(s-1)}, \boldsymbol{y}) p(\theta_1^{(s-1)}, \theta_2^{(s-1)} \mid \boldsymbol{y}) d\theta_1^{(s-1)} d\theta_2^{(s-1)} \\
=&\int p(\theta_2^{(s)} \mid \theta_1^{(s)}, \boldsymbol{y}) p(\theta_1^{(s)} \mid \theta_2^{(s-1)}, \boldsymbol{y}) \int p(\theta_1^{(s-1)}, \theta_2^{(s-1)} \mid \boldsymbol{y}) d\theta_1^{(s-1)} d\theta_2^{(s-1)} \\
=&p(\theta_2^{(s)} \mid \theta_1^{(s)}, \boldsymbol{y}) \int p(\theta_1^{(s)} \mid \theta_2^{(s-1)}, \boldsymbol{y}) p(\theta_2^{(s-1)} \mid \boldsymbol{y}) d\theta_2^{(s-1)} \\
=&p(\theta_2^{(s)} \mid \theta_1^{(s)}, \boldsymbol{y}) p(\theta_1^{(s)} \mid \boldsymbol{y}) \\
=&p(\theta_1^{(s)}, \theta_2^{(s)} \mid \boldsymbol{y})
\end{aligned}$$

即 $\boldsymbol{\theta}^{(s)}$ 也服从后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. □

- **Irreducibility**. 如果从任意初始状态 $\boldsymbol{x}' \in \mathcal{X}$ 出发，Markov process 可以在有限步到达任意其它状态 $\boldsymbol{x} \in \mathcal{X}$，称 Markov process 为 irreducible. 严格的定义如下：

  Definition 3.1 (Irreducibility)**.** 如果 $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $\exists k < \infty$ 使得 k-step transition probability $P^k(\boldsymbol{x} \mid \boldsymbol{x}') > 0$, 称该 Markov process 为 irreducible.

  如果 Markov chain 的 transition density 满足 $p(\boldsymbol{x} \mid \boldsymbol{x}') > 0$, $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, 则该 Markov chain 是 irreducible.

  - A non-irreducible Markov process. 令状态空间 $\mathcal{X} = (-1, 1)$ 上的一个 Markov chain 有如下的 transition distribution：

    $$p(x \mid x') = \begin{cases} x \sim U(0,1), & \text{if } x' \geq 0 \\ x \sim U(-1,0), & \text{if } x' < 0 \end{cases}$$

    该过程不是 irreducible，因为从任意 $x' < 0$ 出发，该过程会一直停留在区间 $(-1, 0)$ 上，无法到达 $\mathcal{X}$ 的另一半区间 $[0, 1)$; 同理，如果从任意 $x' > 0$ 出发，该过程会一直停留在区间 $[0, 1)$ 上。即总有一半的状态空间是可以 “去掉” 的。

- **非周期性（Aperiodicity）**. 如果 $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $\gcd\{k : P^k(\boldsymbol{x} \mid \boldsymbol{x}') > 0\} = 1$, 其中 gcd 表示该整数集合的最大公约数 (greatest common divisor), 称该过程具有**非周期性**。显然当 transition density 满足 $p(\boldsymbol{x} \mid \boldsymbol{x}') > 0$, $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$，Markov chain 是非周期的.

  – 一个周期性的 Markov process. 令状态空间 $\mathcal{X} = (-1, 1)$ 上的一个 Markov chain 有如下的 transition distribution：

  $$p(x \mid x') = \begin{cases} x \sim U(0,1), & \text{if } x' < 0 \\ x \sim U(-1,0), & \text{if } x' \geq 0 \end{cases}$$

  该过程是 irreducible 但不是非周期的：如果该 Markov chain 从 $x' \in A \subseteq (-1, 0)$ 出发，它需要先移动到区间 $[0, 1)$ 上才能再次返回集合 A 上。因此对任何集合 $A \subseteq (-1, 0)$ 或 $A \subseteq [0, 1)$，Markov chain 每经过两步才有可能访问到 A, 此时称 A 的周期为 2.

- **遍历性（Ergodicity）**. 我们称一个 irreducible 且非周期的 Markov chain 具有**遍历性**。具有遍历性的 Markov chain 有以下重要性质：

  1. 存在唯一的 stationary distribution $\pi(\boldsymbol{x})$.
  2. 从任意初始值出发，该过程都会收敛到 stationary distribution $\pi(\boldsymbol{x})$, 即对 $\forall \boldsymbol{x}' \in \mathcal{X}$, 当 $k \to \infty$,

  $$p^k(\boldsymbol{x} \mid \boldsymbol{x}') \to \pi(\boldsymbol{x}).$$

  这说明对遍历的 Markov chain, 从任何初始状态 $\boldsymbol{x}'$ 出发，未来状态 $\boldsymbol{x}$ 的分布会随 k 的增加越来越接近 stationary distribution $\pi(\boldsymbol{x})$. 即初始状态会被 “遗忘”，当 $k \to \infty$, 该过程产生的 $\boldsymbol{x}^{(k)} \sim \pi(\boldsymbol{x})$.

  因此对遍历的 Markov chain $\{\boldsymbol{x}^{(t)}\}$, 样本均值会 (almost surely) 收敛到 stationary distribution 的期望，且对任何可积的函数 $g(\cdot)$, 当 $T \to \infty$,

  $$\frac{1}{T} \sum_{t=1}^{T} g(\boldsymbol{x}^{(t)}) \to \int g(\boldsymbol{x}) \pi(\boldsymbol{x}) d\boldsymbol{x}.$$

  一个 irreducible 且非周期的 Markov chain 等价于 $\exists k < \infty$ 使得 $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $P^k(\boldsymbol{x} \mid \boldsymbol{x}') > 0$. 在更严格的定义中，Markov chain 的遍历性还需要满足 positive recurrence 条件: 从 $\mathcal{X}$ 上的任意集合 A 的一点 $\boldsymbol{x} \in A$ 出发，持续运行 Markov chain，它可以无穷次回到 A；或者 Markov chain 会在期望有限的步数内再次回到 A. 当 transition density 满足 $p(\boldsymbol{x} \mid \boldsymbol{x}') > 0$, $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ 时，irreducibility 与 positive recurrence 是等价的。在有些情形下，irreducible Markov chain 不一定是 positive recurrent.

在 Bayesian 分析中，通过 Gibbs sampler 产生的 Markov chain 一般是遍历的，上述结果表明通过产生一条很长的 Markov chain，该过程的样本可以近似描述后验分布（stationary distribution).

有时当初始值选取得不太好，Markov chain 可能会移动很长时间才收敛到 stationary distribution 概率（密度）较高的区域，我们称这段时间为 *burn-in.* 用样本均值估计目标期望时通常舍弃 burn-in 阶段的样本。

References

Hoff, P. D. (2009). *A first course in Bayesian statistical methods.* Springer Science & Business Media.

