SDS 383C: Statistical Modeling I
Fall 2022, Module VII

Abhra Sarkar

Department of Statistics and Data Sciences
The University of Texas at Austin

"All models are wrong, but some are useful."- George E. P. Box

Monte Carlo Integration

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y})d\mathbf{y}, \quad \mathcal{Y} \subseteq \mathbb{R}^d$.
- Let $h(\mathbf{y}) = g(\mathbf{y})p(\mathbf{y})$ where $p(\mathbf{y})$ is a density on $\mathcal{Y}$. Then

$$\mathbf{I} = \int_{\mathcal{Y}} h(\mathbf{y})d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})} g(\mathbf{y}).$$

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $p(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{p(\mathbf{y})} g(\mathbf{y}) \,\widehat{=}\, \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{y}_i) = \widehat{\mathbf{I}}.$$

- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}) \,\widehat{=}\, \frac{1}{n(n-1)}\sum_{i=1}^{n}\left\{g(\mathbf{y}_i) - \widehat{\mathbf{I}}\right\}^2.$$

- The joint posterior may be complex or known only up to a normalizing constant but we may need to evaluate $\int g(\boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})d\boldsymbol{\theta}$.
- We may still be able to sample from the posterior and evaluate this as

$$\int g(\boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})d\boldsymbol{\theta} \,\widehat{=}\, \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\theta}_i).$$

Monte Carlo Integration

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y}) d\mathbf{y}, \quad \mathcal{Y} \subseteq \mathbb{R}^d$.
- Let $h(\mathbf{y}) = g(\mathbf{y})p(\mathbf{y})$ where $p(\mathbf{y})$ is a density on $\mathcal{Y}$. Then

$$\mathbf{I} = \int_{\mathcal{Y}} h(\mathbf{y}) d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y}) d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} g(\mathbf{y}).$$

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $p(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{p(\mathbf{y})} g(\mathbf{y}) \widehat{=} \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{y}_i) = \widehat{\mathbf{I}}.$$

- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}) \widehat{=} \frac{1}{n(n-1)} \sum_{i=1}^{n} \left\{ g(\mathbf{y}_i) - \widehat{\mathbf{I}} \right\}^2.$$

- The joint posterior may be complex or known only up to a normalizing constant but we may need to evaluate $\int g(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n}) d\boldsymbol{\theta}$.
- We may still be able to sample from the posterior and evaluate this as

$$\int g(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n}) d\boldsymbol{\theta} \widehat{=} \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{\theta}_i).$$

Monte Carlo Integration

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y}) d\mathbf{y}, \quad \mathcal{Y} \subseteq \mathbb{R}^d$.
- Let $h(\mathbf{y}) = g(\mathbf{y}) p(\mathbf{y})$ where $p(\mathbf{y})$ is a density on $\mathcal{Y}$. Then

$$\mathbf{I} = \int_{\mathcal{Y}} h(\mathbf{y}) d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} g(\mathbf{y}).$$

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $p(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{p(\mathbf{y})} g(\mathbf{y}) \widehat{=} \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{y}_i) = \widehat{\mathbf{I}}.$$

- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}) \widehat{=} \frac{1}{n(n-1)} \sum_{i=1}^{n} \left\{ g(\mathbf{y}_i) - \widehat{\mathbf{I}} \right\}^2.$$

- The joint posterior may be complex or known only up to a normalizing constant but we may need to evaluate $\int g(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n}) d\boldsymbol{\theta}$.
- We may still be able to sample from the posterior and evaluate this as

$$\int g(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n}) d\boldsymbol{\theta} \widehat{=} \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{\theta}_i).$$

Monte Carlo Integration

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y})d\mathbf{y}, \quad \mathcal{Y} \subseteq \mathbb{R}^d$.
- Let $h(\mathbf{y}) = g(\mathbf{y})p(\mathbf{y})$ where $p(\mathbf{y})$ is a density on $\mathcal{Y}$. Then

$$\mathbf{I} = \int_{\mathcal{Y}} h(\mathbf{y})d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})} g(\mathbf{y}).$$

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $p(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{p(\mathbf{y})} g(\mathbf{y}) \mathrel{\widehat{=}} \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{y}_i) = \widehat{\mathbf{I}}.$$

- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}) \mathrel{\widehat{=}} \frac{1}{n(n-1)}\sum_{i=1}^{n} \left\{ g(\mathbf{y}_i) - \widehat{\mathbf{I}} \right\}^2 .$$

- The joint posterior may be complex or known only up to a normalizing constant but we may need to evaluate $\int g(\boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})d\boldsymbol{\theta}$.
- We may still be able to sample from the posterior and evaluate this as

$$\int g(\boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})d\boldsymbol{\theta} \mathrel{\widehat{=}} \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\theta}_i).$$

Monte Carlo Integration

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y})d\mathbf{y}, \quad \mathcal{Y} \subseteq \mathbb{R}^d$.
- Let $h(\mathbf{y}) = g(\mathbf{y})p(\mathbf{y})$ where $p(\mathbf{y})$ is a density on $\mathcal{Y}$. Then

$$\mathbf{I} = \int_{\mathcal{Y}} h(\mathbf{y})d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})} g(\mathbf{y}).$$

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $p(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{p(\mathbf{y})} g(\mathbf{y}) \mathrel{\widehat{=}} \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{y}_i) = \widehat{\mathbf{I}}.$$

- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}) \mathrel{\widehat{=}} \frac{1}{n(n-1)}\sum_{i=1}^{n}\left\{g(\mathbf{y}_i) - \widehat{\mathbf{I}}\right\}^2.$$

$$\overbrace{p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})}^{posterior} = \frac{p(\boldsymbol{\theta})p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta})}{p(\mathbf{y}_{1:n})} = \frac{\overbrace{p(\boldsymbol{\theta})}^{prior}\,\overbrace{p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta})}^{likelihood}}{\int \underbrace{p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta})}_{likelihood}\,\underbrace{p(\boldsymbol{\theta})}_{prior}\, d\boldsymbol{\theta}} \propto \overbrace{p(\boldsymbol{\theta})}^{prior}\,\overbrace{p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta})}^{likelihood}$$

- The joint posterior may be complex or known only up to a normalizing constant but we may need to evaluate $\int g(\boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})d\boldsymbol{\theta}$.
- We may still be able to sample from the posterior and evaluate this as

$$\int g(\boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})d\boldsymbol{\theta} \mathrel{\widehat{=}} \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\theta}_i).$$

Monte Carlo Integration

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y})d\mathbf{y}, \quad \mathcal{Y} \subseteq \mathbb{R}^d$.
- Let $h(\mathbf{y}) = g(\mathbf{y})p(\mathbf{y})$ where $p(\mathbf{y})$ is a density on $\mathcal{Y}$. Then

$$\mathbf{I} = \int_{\mathcal{Y}} h(\mathbf{y})d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})} g(\mathbf{y}).$$

- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $p(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{p(\mathbf{y})} g(\mathbf{y}) \mathrel{\widehat{=}} \tfrac{1}{n} \textstyle\sum_{i=1}^n g(\mathbf{y}_i) = \widehat{\mathbf{I}}.$$

- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}) \mathrel{\widehat{=}} \tfrac{1}{n(n-1)} \textstyle\sum_{i=1}^n \left\{ g(\mathbf{y}_i) - \widehat{\mathbf{I}} \right\}^2 .$$

$$\overbrace{p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})}^{posterior} = \frac{p(\boldsymbol{\theta})p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta})}{p(\mathbf{y}_{1:n})} = \frac{\overbrace{p(\boldsymbol{\theta})}^{prior}\,\overbrace{p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta})}^{likelihood}}{\int \underbrace{p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta})}_{likelihood}\,\underbrace{p(\boldsymbol{\theta})}_{prior}\, d\boldsymbol{\theta}} \propto \overbrace{p(\boldsymbol{\theta})}^{prior}\,\overbrace{p(\mathbf{y}_{1:n} \mid \boldsymbol{\theta})}^{likelihood}$$

- The joint posterior may be complex or known only up to a normalizing constant but we may need to evaluate $\int g(\boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})d\boldsymbol{\theta}$.
- We may still be able to sample from the posterior and evaluate this as

$$\int g(\boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})d\boldsymbol{\theta} \mathrel{\widehat{=}} \tfrac{1}{n} \textstyle\sum_{i=1}^n g(\boldsymbol{\theta}_i).$$

Importance Sampling

- The density $p(\mathbf{y})$ is of interest but is difficult to sample from.
- The density $q(\mathbf{y})$ roughly approximates $p(\mathbf{y})$ and is easier to sample from.
- Then $\mathbf{I} = \int_{\mathcal{Y}} g(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} q(\mathbf{y}) d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\}$.
- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\} \mathrel{\widehat{=}} \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{y}_i) \frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)} = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{y}_i) w_i = \widehat{\mathbf{I}}_1.$$

- An alternative formulation is

$$\mathbf{I} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\} \mathrel{\widehat{=}} \frac{\sum_{i=1}^{n} w_i g(\mathbf{y}_i)}{\sum_{i=1}^{n} w_i} = \widehat{\mathbf{I}}_2 \quad \text{where} \quad w_i = \frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)}.$$

- $\mathbb{E}(w_i) = \int \frac{p(\mathbf{y})}{q(\mathbf{y})} q(\mathbf{y}) d\mathbf{y} = \int p(\mathbf{y}) d\mathbf{y} = 1$ so that $\mathbb{E}\left(\sum_{i=1}^{n} w_i\right) = n$.
- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}_2) \mathrel{\widehat{=}} \frac{\sum_{i=1}^{n} w_i \left\{ g(\mathbf{y}_i) - \widehat{\mathbf{I}}_2 \right\}^2}{\sum_{i=1}^{n} w_i}.$$

- $\widehat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\widehat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \to \infty$.
- $\widehat{\mathbf{I}}_2$ also works when the normalizing constant of $p(\mathbf{y})$ is NOT known.

Importance Sampling

- The density $p(\mathbf{y})$ is of interest but is difficult to sample from.
- The density $q(\mathbf{y})$ roughly approximates $p(\mathbf{y})$ and is easier to sample from.
- Then $\mathbf{I} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\}$.
- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \mathrel{\widehat{=}} \frac{1}{n}\sum_{i=1}^n g(\mathbf{y}_i)\frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)} = \frac{1}{n}\sum_{i=1}^n g(\mathbf{y}_i)w_i = \widehat{\mathbf{I}}_1.$$

- An alternative formulation is

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \mathrel{\widehat{=}} \frac{\sum_{i=1}^n w_i g(\mathbf{y}_i)}{\sum_{i=1}^n w_i} = \widehat{\mathbf{I}}_2 \quad \text{where} \quad w_i = \frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)}.$$

- $\mathbb{E}(w_i) = \int \frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \int p(\mathbf{y})d\mathbf{y} = 1$ so that $\mathbb{E}\left(\sum_{i=1}^n w_i\right) = n$.
- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}_2) \mathrel{\widehat{=}} \frac{\sum_{i=1}^n w_i\left\{g(\mathbf{y}_i) - \widehat{\mathbf{I}}_2\right\}^2}{\sum_{i=1}^n w_i}.$$

- $\widehat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\widehat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \to \infty$.
- $\widehat{\mathbf{I}}_2$ also works when the normalizing constant of $p(\mathbf{y})$ is NOT known.

Importance Sampling

- The density $p(\mathbf{y})$ is of interest but is difficult to sample from.
- The density $q(\mathbf{y})$ roughly approximates $p(\mathbf{y})$ and is easier to sample from.
- Then $\mathbf{I} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\}.$
- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \widehat{=} \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{y}_i)\frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)} = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{y}_i)w_i = \widehat{\mathbf{I}}_1.$$

- An alternative formulation is

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \widehat{=} \frac{\sum_{i=1}^{n} w_i g(\mathbf{y}_i)}{\sum_{i=1}^{n} w_i} = \widehat{\mathbf{I}}_2 \quad \text{where} \quad w_i = \frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)}.$$

- $\mathbb{E}(w_i) = \int \frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \int p(\mathbf{y})d\mathbf{y} = 1$ so that $\mathbb{E}\left(\sum_{i=1}^{n} w_i\right) = n$.
- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}_2) \widehat{=} \frac{\sum_{i=1}^{n} w_i\left\{g(\mathbf{y}_i) - \widehat{\mathbf{I}}_2\right\}^2}{\sum_{i=1}^{n} w_i}.$$

- $\widehat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\widehat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \to \infty$.
- $\widehat{\mathbf{I}}_2$ also works when the normalizing constant of $p(\mathbf{y})$ is NOT known.

Importance Sampling

- The density $p(\mathbf{y})$ is of interest but is difficult to sample from.
- The density $q(\mathbf{y})$ roughly approximates $p(\mathbf{y})$ and is easier to sample from.
- Then $\mathbf{I} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\}$.
- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \mathrel{\widehat{=}} \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{y}_i)\frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)} = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{y}_i)w_i = \widehat{\mathbf{I}}_1.$$

- An alternative formulation is

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \mathrel{\widehat{=}} \frac{\sum_{i=1}^{n} w_i g(\mathbf{y}_i)}{\sum_{i=1}^{n} w_i} = \widehat{\mathbf{I}}_2 \quad \text{where} \quad w_i = \frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)}.$$

- $\mathbb{E}(w_i) = \int \frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \int p(\mathbf{y})d\mathbf{y} = 1$ so that $\mathbb{E}\left(\sum_{i=1}^{n} w_i\right) = n$.
- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}_2) \mathrel{\widehat{=}} \frac{\sum_{i=1}^{n} w_i\left\{g(\mathbf{y}_i)-\widehat{\mathbf{I}}_2\right\}^2}{\sum_{i=1}^{n} w_i}.$$

- $\widehat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\widehat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \to \infty$.
- $\widehat{\mathbf{I}}_2$ also works when the normalizing constant of $p(\mathbf{y})$ is NOT known.

Importance Sampling

- The density $p(\mathbf{y})$ is of interest but is difficult to sample from.
- The density $q(\mathbf{y})$ roughly approximates $p(\mathbf{y})$ and is easier to sample from.
- Then $\mathbf{I} = \int_{\mathcal{Y}} g(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} q(\mathbf{y}) d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\}$.
- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\} \widehat{=} \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{y}_i) \frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)} = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{y}_i) w_i = \widehat{\mathbf{I}}_1 .$$

- An alternative formulation is

$$\mathbf{I} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\} \widehat{=} \frac{\sum_{i=1}^{n} w_i g(\mathbf{y}_i)}{\sum_{i=1}^{n} w_i} = \widehat{\mathbf{I}}_2 \quad \text{where} \quad w_i = \frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)}.$$

- $\mathbb{E}(w_i) = \int \frac{p(\mathbf{y})}{q(\mathbf{y})} q(\mathbf{y}) d\mathbf{y} = \int p(\mathbf{y}) d\mathbf{y} = 1$ so that $\mathbb{E}\left(\sum_{i=1}^{n} w_i\right) = n$.
- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}_2) \widehat{=} \frac{\sum_{i=1}^{n} w_i \left\{ g(\mathbf{y}_i) - \widehat{\mathbf{I}}_2 \right\}^2}{\sum_{i=1}^{n} w_i}.$$

- $\widehat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\widehat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \to \infty$.
- $\widehat{\mathbf{I}}_2$ also works when the normalizing constant of $p(\mathbf{y})$ is NOT known.

Importance Sampling

- The density $p(\mathbf{y})$ is of interest but is difficult to sample from.
- The density $q(\mathbf{y})$ roughly approximates $p(\mathbf{y})$ and is easier to sample from.
- Then $\mathbf{I} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\}$.
- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \widehat{=} \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{y}_i)\frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)} = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{y}_i)w_i = \widehat{\mathbf{I}}_1.$$

- An alternative formulation is

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \widehat{=} \frac{\sum_{i=1}^{n} w_i g(\mathbf{y}_i)}{\sum_{i=1}^{n} w_i} = \widehat{\mathbf{I}}_2 \quad \text{where} \quad w_i = \frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)}.$$

- $\mathbb{E}(w_i) = \int \frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \int p(\mathbf{y})d\mathbf{y} = 1$ so that $\mathbb{E}\left(\sum_{i=1}^{n} w_i\right) = n$.
- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}_2) \widehat{=} \frac{\sum_{i=1}^{n} w_i\left\{g(\mathbf{y}_i) - \widehat{\mathbf{I}}_2\right\}^2}{\sum_{i=1}^{n} w_i}.$$

- $\widehat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\widehat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \to \infty$.
- $\widehat{\mathbf{I}}_2$ also works when the normalizing constant of $p(\mathbf{y})$ is NOT known.

Importance Sampling

- The density $p(\mathbf{y})$ is of interest but is difficult to sample from.
- The density $q(\mathbf{y})$ roughly approximates $p(\mathbf{y})$ and is easier to sample from.
- Then $\mathbf{I} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\}.$
- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \widehat{=} \frac{1}{n}\sum_{i=1}^n g(\mathbf{y}_i)\frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)} = \frac{1}{n}\sum_{i=1}^n g(\mathbf{y}_i)w_i = \widehat{\mathbf{I}}_1.$$

- An alternative formulation is

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \widehat{=} \frac{\sum_{i=1}^n w_i g(\mathbf{y}_i)}{\sum_{i=1}^n w_i} = \widehat{\mathbf{I}}_2 \quad \text{where} \quad w_i = \frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)}.$$

- $\mathbb{E}(w_i) = \int \frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \int p(\mathbf{y})d\mathbf{y} = 1$ so that $\mathbb{E}\left(\sum_{i=1}^n w_i\right) = n.$
- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}_2) \widehat{=} \frac{\sum_{i=1}^n w_i\left\{g(\mathbf{y}_i)-\widehat{\mathbf{I}}_2\right\}^2}{\sum_{i=1}^n w_i}.$$

- $\widehat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\widehat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \to \infty$.
- $\widehat{\mathbf{I}}_2$ also works when the normalizing constant of $p(\mathbf{y})$ is NOT known.

- The density $p(\mathbf{y})$ is of interest but is difficult to sample from.
- The density $q(\mathbf{y})$ roughly approximates $p(\mathbf{y})$ and is easier to sample from.
- Then $\mathbf{I} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\}$.
- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \,\widehat{=}\, \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{y}_i)\frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)} = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{y}_i)w_i = \widehat{\mathbf{I}}_1.$$

- An alternative formulation is

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \,\widehat{=}\, \frac{\sum_{i=1}^{n} w_i g(\mathbf{y}_i)}{\sum_{i=1}^{n} w_i} = \widehat{\mathbf{I}}_2 \quad \text{where} \quad w_i = \frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)}.$$

- $\mathbb{E}(w_i) = \int \frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \int p(\mathbf{y})d\mathbf{y} = 1$ so that $\mathbb{E}\left(\sum_{i=1}^{n} w_i\right) = n$.
- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}_2) \,\widehat{=}\, \frac{\sum_{i=1}^{n} w_i\left\{g(\mathbf{y}_i)-\widehat{\mathbf{I}}_2\right\}^2}{\sum_{i=1}^{n} w_i}.$$

- $\widehat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\widehat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \to \infty$.
- $\widehat{\mathbf{I}}_2$ also works when the normalizing constant of $p(\mathbf{y})$ is NOT known.

- The density $p(\mathbf{y})$ is of interest but is difficult to sample from.
- The density $q(\mathbf{y})$ roughly approximates $p(\mathbf{y})$ and is easier to sample from.
- Then $\mathbf{I} = \int_{\mathcal{Y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathcal{Y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\}$.
- Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \widehat{=} \frac{1}{n}\sum_{i=1}^n g(\mathbf{y}_i)\frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)} = \frac{1}{n}\sum_{i=1}^n g(\mathbf{y}_i)w_i = \widehat{\mathbf{I}}_1.$$

- An alternative formulation is

$$\mathbf{I} = \mathbb{E}_{\mathbf{y}\sim q(\mathbf{y})}\left\{g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}\right\} \widehat{=} \frac{\sum_{i=1}^n w_i g(\mathbf{y}_i)}{\sum_{i=1}^n w_i} = \widehat{\mathbf{I}}_2 \quad \text{where} \quad w_i = \frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)}.$$

- $\mathbb{E}(w_i) = \int \frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \int p(\mathbf{y})d\mathbf{y} = 1$ so that $\mathbb{E}\left(\sum_{i=1}^n w_i\right) = n$.
- Monte Carlo standard error

$$SE^2(\widehat{\mathbf{I}}_2) \widehat{=} \frac{\sum_{i=1}^n w_i\left\{g(\mathbf{y}_i) - \widehat{\mathbf{I}}_2\right\}^2}{\sum_{i=1}^n w_i}.$$

- $\widehat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\widehat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \to \infty$.
- $\widehat{\mathbf{I}}_2$ also works when the normalizing constant of $p(\mathbf{y})$ is NOT known.

Importance Sampling

- Likelihood: $y_1, \ldots, y_n \overset{iid}{\sim}$ Bernoulli(θ) with $p(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$
- Prior: $p(\theta) \propto \cos^2(4\pi\theta) = \widetilde{p}(\theta)$
- Posterior: $p(\theta \mid \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta)\, \theta^s(1-\theta)^{n-s} = \widetilde{p}(\theta \mid \mathbf{y}_{1:n})$

Importance Sampling

- Likelihood: $y_1, \ldots, y_n \overset{iid}{\sim}$ Bernoulli(θ) with $p(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$
- Prior: $p(\theta) \propto \cos^2(4\pi\theta) = \widetilde{p}(\theta)$
- Posterior: $p(\theta \mid \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta)\, \theta^s(1-\theta)^{n-s} = \widetilde{p}(\theta \mid \mathbf{y}_{1:n})$

$$p(\theta) = \frac{\widetilde{p}(\theta)}{\mathbf{I}_{prior}}, \qquad p(\theta \mid \mathbf{y}_{1:n}) = \frac{\widetilde{p}(\theta|\mathbf{y}_{1:n})}{\mathbf{I}_{post}}, \quad \text{where}$$

$$\mathbf{I}_{prior} = \int_\Theta \widetilde{p}(\theta) d\theta = \int_\Theta \frac{\widetilde{p}(\theta)}{q(\theta)} q(\theta) d\theta = \mathbb{E}_{\theta \sim q(\theta)} \left\{ \frac{\widetilde{p}(\theta)}{q(\theta)} \right\},$$

$$\mathbf{I}_{post} = \int_\Theta \widetilde{p}(\theta \mid \mathbf{y}_{1:n}) d\theta = \int_\Theta \frac{\widetilde{p}(\theta|\mathbf{y}_{1:n})}{q(\theta)} q(\theta) d\theta = \mathbb{E}_{\theta \sim q(\theta)} \left\{ \frac{\widetilde{p}(\theta|\mathbf{y}_{1:n})}{q(\theta)} \right\}.$$

Importance Sampling

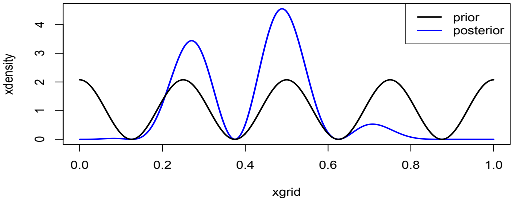
- Likelihood: $y_1, \ldots, y_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$ with $p(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$
- Prior: $p(\theta) \propto \cos^2(4\pi\theta) = \widetilde{p}(\theta)$
- Posterior: $p(\theta \mid \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta)\, \theta^s(1-\theta)^{n-s} = \widetilde{p}(\theta \mid \mathbf{y}_{1:n})$

$$p(\theta) = \frac{\widetilde{p}(\theta)}{\mathbf{I}_{prior}}, \qquad p(\theta \mid \mathbf{y}_{1:n}) = \frac{\widetilde{p}(\theta|\mathbf{y}_{1:n})}{\mathbf{I}_{post}}, \quad \text{where}$$

$$\mathbf{I}_{prior} = \int_\Theta \widetilde{p}(\theta) d\theta = \int_\Theta \frac{\widetilde{p}(\theta)}{q(\theta)} q(\theta) d\theta = \mathbb{E}_{\theta\sim q(\theta)}\left\{\frac{\widetilde{p}(\theta)}{q(\theta)}\right\},$$

$$\mathbf{I}_{post} = \int_\Theta \widetilde{p}(\theta \mid \mathbf{y}_{1:n}) d\theta = \int_\Theta \frac{\widetilde{p}(\theta|\mathbf{y}_{1:n})}{q(\theta)} q(\theta) d\theta = \mathbb{E}_{\theta\sim q(\theta)}\left\{\frac{\widetilde{p}(\theta|\mathbf{y}_{1:n})}{q(\theta)}\right\}.$$

- Importance sampling density: $q(\theta) = \text{Beta}(2,2)$
- Iterations $M = 500,000$
- $\mathbf{I}_{prior} \approx 0.4940351$
- $n = 10,\ s = \sum_i y_i = 4$
- $\mathbf{I}_{post} \approx 0.0002172$



Importance Resampling

- The density $p(y)$ is of interest but is difficult to sample from.
- The density $q(y)$ roughly approximates $p(y)$ and is easier to sample from.
- Then

$$F_p(y) = \int_{-\infty}^{y} p(z)dz = \int_{-\infty}^{\infty} 1(z \le y)p(z)dz = \int_{-\infty}^{\infty} 1(z \le y)p(z)dz$$
$$= \int_{-\infty}^{\infty} 1(z \le y)\frac{p(z)}{q(z)}q(z)dz = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\frac{p(z)}{q(z)}\right\}.$$

- Let $z_1, \ldots, z_n$ be a random sample from $q(z)$. Then

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\frac{p(z)}{q(z)}\right\} \mathrel{\hat{=}} \frac{1}{n}\sum_{i=1}^{n} 1(z_i \le y)\frac{p(z_i)}{q(z_i)} = \frac{1}{n}\sum_{i=1}^{n} 1(z_i \le y)w_i = \widehat{F}_1(y).$$

- An alternative formulation is

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\frac{p(z)}{q(z)}\right\} \mathrel{\hat{=}} \frac{\sum_{i=1}^{n} w_i 1(z_i \le y)}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} \widetilde{w}_i 1(z_i \le y) = \widehat{F}_2(y).$$

- $\widehat{F}_1(y) \xrightarrow{P} F_p(y)$ and $\widehat{F}_2(y) \xrightarrow{P} F_p(y)$ as $n \to \infty$.
- $\widehat{F}_2(y)$ also works when the normalizing constant of $p(z)$ is NOT known.
- $\widehat{F}_2(y)$ shows that sampling from $p(y)$ can be approximated by resampling the z_i's with weights $\widetilde{w}_i$'s.

Importance Resampling

- The density $p(y)$ is of interest but is difficult to sample from.
- The density $q(y)$ roughly approximates $p(y)$ and is easier to sample from.
- Then

$$F_p(y) = \int_{-\infty}^{y} p(z)dz = \int_{-\infty}^{\infty} 1(z \leq y)p(z)dz = \int_{-\infty}^{\infty} 1(z \leq y)p(z)dz$$
$$= \int_{-\infty}^{\infty} 1(z \leq y)\frac{p(z)}{q(z)}q(z)dz = \mathbb{E}_{z\sim q(z)}\left\{1(z \leq y)\frac{p(z)}{q(z)}\right\}.$$

- Let $z_1, \ldots, z_n$ be a random sample from $q(z)$. Then

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \leq y)\frac{p(z)}{q(z)}\right\} \widehat{=} \frac{1}{n}\sum_{i=1}^{n} 1(z_i \leq y)\frac{p(z_i)}{q(z_i)} = \frac{1}{n}\sum_{i=1}^{n} 1(z_i \leq y)w_i = \widehat{F}_1(y).$$

- An alternative formulation is

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \leq y)\frac{p(z)}{q(z)}\right\} \widehat{=} \frac{\sum_{i=1}^{n} w_i 1(z_i \leq y)}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} \widetilde{w}_i 1(z_i \leq y) = \widehat{F}_2(y).$$

- $\widehat{F}_1(y) \xrightarrow{P} F_p(y)$ and $\widehat{F}_2(y) \xrightarrow{P} F_p(y)$ as $n \to \infty$.
- $\widehat{F}_2(y)$ also works when the normalizing constant of $p(z)$ is NOT known.
- $\widehat{F}_2(y)$ shows that sampling from $p(y)$ can be approximated by resampling the z_i's with weights $\widetilde{w}_i$'s.

Importance Resampling

- The density $p(y)$ is of interest but is difficult to sample from.
- The density $q(y)$ roughly approximates $p(y)$ and is easier to sample from.
- Then

$$\begin{aligned} F_p(y) &= \int_{-\infty}^{y} p(z)dz = \int_{-\infty}^{\infty} 1(z \le y)p(z)dz = \int_{-\infty}^{\infty} 1(z \le y)p(z)dz \\ &= \int_{-\infty}^{\infty} 1(z \le y)\frac{p(z)}{q(z)}q(z)dz = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\frac{p(z)}{q(z)}\right\}. \end{aligned}$$

- Let $z_1, \ldots, z_n$ be a random sample from $q(z)$. Then

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\frac{p(z)}{q(z)}\right\} \widehat{=} \frac{1}{n}\sum_{i=1}^{n} 1(z_i \le y)\frac{p(z_i)}{q(z_i)} = \frac{1}{n}\sum_{i=1}^{n} 1(z_i \le y)w_i = \widehat{F}_1(y).$$

- An alternative formulation is

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\frac{p(z)}{q(z)}\right\} \widehat{=} \frac{\sum_{i=1}^{n} w_i 1(z_i \le y)}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} \widetilde{w}_i 1(z_i \le y) = \widehat{F}_2(y).$$

- $\widehat{F}_1(y) \xrightarrow{P} F_p(y)$ and $\widehat{F}_2(y) \xrightarrow{P} F_p(y)$ as $n \to \infty$.
- $\widehat{F}_2(y)$ also works when the normalizing constant of $p(z)$ is NOT known.
- $\widehat{F}_2(y)$ shows that sampling from $p(y)$ can be approximated by resampling the z_i's with weights $\widetilde{w}_i$'s.

Importance Resampling

- The density $p(y)$ is of interest but is difficult to sample from.
- The density $q(y)$ roughly approximates $p(y)$ and is easier to sample from.
- Then

$$\begin{aligned} F_p(y) &= \int_{-\infty}^{y} p(z)dz = \int_{-\infty}^{\infty} 1(z \le y)p(z)dz = \int_{-\infty}^{\infty} 1(z \le y)p(z)dz \\ &= \int_{-\infty}^{\infty} 1(z \le y)\tfrac{p(z)}{q(z)}q(z)dz = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\tfrac{p(z)}{q(z)}\right\}. \end{aligned}$$

- Let $z_1, \ldots, z_n$ be a random sample from $q(z)$. Then

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\tfrac{p(z)}{q(z)}\right\} \;\widehat{=}\; \tfrac{1}{n}\sum_{i=1}^{n} 1(z_i \le y)\tfrac{p(z_i)}{q(z_i)} = \tfrac{1}{n}\sum_{i=1}^{n} 1(z_i \le y)w_i = \widehat{F}_1(y).$$

- An alternative formulation is

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\tfrac{p(z)}{q(z)}\right\} \;\widehat{=}\; \frac{\sum_{i=1}^{n} w_i 1(z_i \le y)}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} \widetilde{w}_i 1(z_i \le y) = \widehat{F}_2(y).$$

- $\widehat{F}_1(y) \xrightarrow{P} F_p(y)$ and $\widehat{F}_2(y) \xrightarrow{P} F_p(y)$ as $n \to \infty$.
- $\widehat{F}_2(y)$ also works when the normalizing constant of $p(z)$ is NOT known.
- $\widehat{F}_2(y)$ shows that sampling from $p(y)$ can be approximated by resampling the z_i's with weights $\widetilde{w}_i$'s.

Importance Resampling

- The density $p(y)$ is of interest but is difficult to sample from.
- The density $q(y)$ roughly approximates $p(y)$ and is easier to sample from.
- Then

$$F_p(y) = \int_{-\infty}^{y} p(z)dz = \int_{-\infty}^{\infty} 1(z \le y)p(z)dz = \int_{-\infty}^{\infty} 1(z \le y)p(z)dz$$
$$= \int_{-\infty}^{\infty} 1(z \le y)\tfrac{p(z)}{q(z)}q(z)dz = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\tfrac{p(z)}{q(z)}\right\}.$$

- Let $z_1, \ldots, z_n$ be a random sample from $q(z)$. Then

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\tfrac{p(z)}{q(z)}\right\} \mathrel{\widehat{=}} \tfrac{1}{n}\textstyle\sum_{i=1}^{n} 1(z_i \le y)\tfrac{p(z_i)}{q(z_i)} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} 1(z_i \le y)w_i = \widehat{F}_1(y).$$

- An alternative formulation is

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \le y)\tfrac{p(z)}{q(z)}\right\} \mathrel{\widehat{=}} \frac{\sum_{i=1}^{n} w_i 1(z_i \le y)}{\sum_{i=1}^{n} w_i} = \textstyle\sum_{i=1}^{n} \widetilde{w}_i 1(z_i \le y) = \widehat{F}_2(y).$$

- $\widehat{F}_1(y) \xrightarrow{P} F_p(y)$ and $\widehat{F}_2(y) \xrightarrow{P} F_p(y)$ as $n \to \infty$.
- $\widehat{F}_2(y)$ also works when the normalizing constant of $p(z)$ is NOT known.
- $\widehat{F}_2(y)$ shows that sampling from $p(y)$ can be approximated by resampling the z_i's with weights $\widetilde{w}_i$'s.

Importance Resampling

- The density $p(y)$ is of interest but is difficult to sample from.
- The density $q(y)$ roughly approximates $p(y)$ and is easier to sample from.
- Then

$$F_p(y) = \int_{-\infty}^{y} p(z)dz = \int_{-\infty}^{\infty} 1(z \leq y)p(z)dz = \int_{-\infty}^{\infty} 1(z \leq y)p(z)dz$$
$$= \int_{-\infty}^{\infty} 1(z \leq y)\frac{p(z)}{q(z)}q(z)dz = \mathbb{E}_{z\sim q(z)}\left\{1(z \leq y)\frac{p(z)}{q(z)}\right\}.$$

- Let $z_1, \ldots, z_n$ be a random sample from $q(z)$. Then

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \leq y)\frac{p(z)}{q(z)}\right\} \mathrel{\widehat{=}} \frac{1}{n}\sum_{i=1}^{n} 1(z_i \leq y)\frac{p(z_i)}{q(z_i)} = \frac{1}{n}\sum_{i=1}^{n} 1(z_i \leq y)w_i = \widehat{F}_1(y).$$

- An alternative formulation is

$$F_p(y) = \mathbb{E}_{z\sim q(z)}\left\{1(z \leq y)\frac{p(z)}{q(z)}\right\} \mathrel{\widehat{=}} \frac{\sum_{i=1}^{n} w_i 1(z_i \leq y)}{\sum_{i=1}^{n} w_i} = \sum_{i=1}^{n} \widetilde{w}_i 1(z_i \leq y) = \widehat{F}_2(y).$$

- $\widehat{F}_1(y) \xrightarrow{P} F_p(y)$ and $\widehat{F}_2(y) \xrightarrow{P} F_p(y)$ as $n \to \infty$.
- $\widehat{F}_2(y)$ also works when the normalizing constant of $p(z)$ is NOT known.
- $\widehat{F}_2(y)$ shows that sampling from $p(y)$ can be approximated by resampling the z_i's with weights $\widetilde{w}_i$'s.

Markov Chains

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \to y_{t-2} \to y_{t-1} \to y_t \to y_{t+1} \to \cdots$$

- **Initial Distribution:** probability that the chain starts with a state y:

$$p(y_0 = y) = \pi_0(y).$$

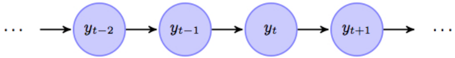
- **Transition Probabilities:** probability that the chain moves to a state y from a state x in a single step:

$$p(y_{t+1} = y \mid y_t = x) = p(y \mid x) = p(x \to y) = p(x, y).$$

- **Transition Probability Matrix:** $\mathbf{P} = \Big(\big(p(x, y)\big)\Big)$.
- The probability that the chain is in state y at time t: $p(y_t = y) = \pi_t(y)$.

Markov Chains

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$$



- **Initial Distribution:** probability that the chain starts with a state y:

$$p(y_0 = y) = \pi_0(y).$$

- **Transition Probabilities:** probability that the chain moves to a state y from a state x in a single step:

$$p(y_{t+1} = y \mid y_t = x) = p(y \mid x) = p(x \to y) = p(x, y).$$

- **Transition Probability Matrix:** $\mathbf{P} = \Big(\Big(p(x,y)\Big)\Big)$.
- The probability that the chain is in state y at time t: $p(y_t = y) = \pi_t(y)$.

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \to y_{t-2} \to y_{t-1} \to y_t \to y_{t+1} \to \cdots$$

- **Initial Distribution:** probability that the chain starts with a state y:

$$p(y_0 = y) = \pi_0(y).$$

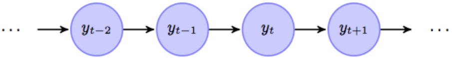
- **Transition Probabilities:** probability that the chain moves to a state y from a state x in a single step:

$$p(y_{t+1} = y \mid y_t = x) = p(y \mid x) = p(x \to y) = p(x, y).$$

- **Transition Probability Matrix: $\mathbf{P} = \Big(\big(p(x, y)\big)\Big).$**
- The probability that the chain is in state y at time t: $p(y_t = y) = \pi_t(y)$.

Markov Chains

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$$



- **Initial Distribution:** probability that the chain starts with a state y:

$$p(y_0 = y) = \pi_0(y).$$

- **Transition Probabilities:** probability that the chain moves to a state y from a state x in a single step:

$$p(y_{t+1} = y \mid y_t = x) = p(y \mid x) = p(x \to y) = p(x, y).$$

- **Transition Probability Matrix:** $\mathbf{P} = \Big(\big(p(x, y)\big)\Big)$.
- The probability that the chain is in state y at time t: $p(y_t = y) = \pi_t(y)$.

Markov Chains

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \to y_{t-2} \to y_{t-1} \to y_t \to y_{t+1} \to \cdots$$

- **Initial Distribution:** probability that the chain starts with a state y:

$$p(y_0 = y) = \pi_0(y).$$

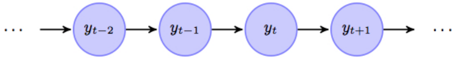
- **Transition Probabilities:** probability that the chain moves to a state y from a state x in a single step:

$$p(y_{t+1} = y \mid y_t = x) = p(y \mid x) = p(x \to y) = p(x, y).$$

- **Transition Probability Matrix:** $\mathbf{P} = \Big(\Big(p(x, y)\Big)\Big).$
- The probability that the chain is in state y at time t: $p(y_t = y) = \pi_t(y)$.
- We have $\sum_{y \in \mathcal{Y}} p(x, y) = 1 \quad \forall\, x \in \mathcal{Y}, \qquad \sum_{y \in \mathcal{Y}} p(y_t = y) = 1.$

Markov Chains

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$$



- **Initial Distribution:** probability that the chain starts with a state y:

$$p(y_0 = y) = \pi_0(y).$$

- **Transition Probabilities:** probability that the chain moves to a state y from a state x in a single step:

$$p(y_{t+1} = y \mid y_t = x) = p(y \mid x) = p(x \to y) = p(x, y).$$

- **Transition Probability Matrix:** $\mathbf{P} = \Big(\big(p(x, y)\big)\Big)$.
- The probability that the chain is in state y at time t: $p(y_t = y) = \pi_t(y)$.
- We have $\quad \mathbf{P}\mathbf{1} = \mathbf{1}, \qquad \boldsymbol{\pi}_t \mathbf{1} = 1.$

Markov Chains

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \to y_{t-2} \to y_{t-1} \to y_t \to y_{t+1} \to \cdots$$

- **Initial Distribution:** probability that the chain starts with a state y:

$$p(y_0 = y) = \pi_0(y).$$

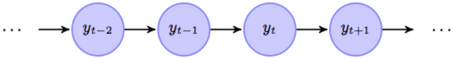
- **Transition Probabilities:** probability that the chain moves to a state y from a state x in a single step:

$$p(y_{t+1} = y \mid y_t = x) = p(y \mid x) = p(x \to y) = p(x, y).$$

- **Transition Probability Matrix:** $\mathbf{P} = \Big(\big(p(x, y)\big)\Big).$
- The probability that the chain is in state y at time t: $p(y_t = y) = \pi_t(y)$.
- We have $\quad \mathbf{P1} = \mathbf{1}, \qquad \boldsymbol{\pi}_t \mathbf{1} = 1.$
- **Chapman-Kolmogorov Equation:** $\pi_{t+1}(y) = p(y_{t+1} = y)$
$$= \textstyle\sum_{x \in \mathcal{Y}} p(y_{t+1} = y \mid y_t = x) p(y_t = x)$$
$$= \textstyle\sum_{x \in \mathcal{Y}} p(x, y) \pi_t(x).$$

Markov Chains

$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$



- **Initial Distribution:** probability that the chain starts with a state y:

$$p(y_0 = y) = \pi_0(y).$$

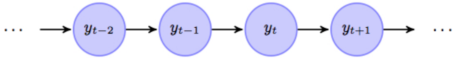
- **Transition Probabilities:** probability that the chain moves to a state y from a state x in a single step:

$$p(y_{t+1} = y \mid y_t = x) = p(y \mid x) = p(x \to y) = p(x, y).$$

- **Transition Probability Matrix:** $\mathbf{P} = \Big(\Big(p(x,y)\Big)\Big)$.
- The probability that the chain is in state y at time t: $p(y_t = y) = \pi_t(y)$.
- We have $\quad \mathbf{P1} = \mathbf{1}, \qquad \boldsymbol{\pi}_t \mathbf{1} = 1.$
- **Chapman-Kolmogorov Equation:** $\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t \mathbf{P}$.

Markov Chains

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$$



- **Initial Distribution:** probability that the chain starts with a state y:

$$p(y_0 = y) = \pi_0(y).$$

- **Transition Probabilities:** probability that the chain moves to a state y from a state x in a single step:

$$p(y_{t+1} = y \mid y_t = x) = p(y \mid x) = p(x \to y) = p(x, y).$$

- **Transition Probability Matrix:** $\mathbf{P} = \Big(\big(p(x, y)\big)\Big)$.
- The probability that the chain is in state y at time t: $p(y_t = y) = \pi_t(y)$.
- We have $\quad \mathbf{P}\mathbf{1} = \mathbf{1}, \qquad \boldsymbol{\pi}_t\mathbf{1} = 1.$
- **Chapman-Kolmogorov Equation:** $\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t\mathbf{P}$.
- We have $\qquad \boldsymbol{\pi}_t = \boldsymbol{\pi}_{t-1}\mathbf{P} = \boldsymbol{\pi}_{t-2}\mathbf{P}^2 = \cdots = \boldsymbol{\pi}_0\mathbf{P}^t.$

Markov Chains - Stationarity

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \to y_{t-2} \to y_{t-1} \to y_t \to y_{t+1} \to \cdots$$

- **Irreducibility:** A Markov chain is irreducible if the chain can move from any state to any other in finite steps - there exists some $n \in \mathbb{N}$ such that

$$p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}.$$

- **Aperiodicity:** A Markov chain is aperiodic if the number of steps required to move between two states is not a multiple of some integer:

$$GCD\{n : p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}\} = 1.$$

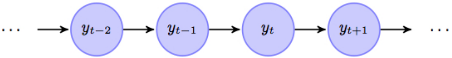
- Aperiodic Markov chains are NOT forced into cycles of fixed lengths between certain states.
- **Stationarity:** A Markov chain may reach stationarity when the probability of being in any particular state is independent of the initial state.
- The stationary distribution π satisfies

$$\pi = \pi \mathbf{P}.$$

- Irreducible and aperiodic Markov chains converge to stationarity.

Markov Chains - Stationarity

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$$



- **Irreducibility:** A Markov chain is irreducible if the chain can move from any state to any other in finite steps - there exists some $n \in \mathbb{N}$ such that

$$p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}.$$

- **Aperiodicity:** A Markov chain is aperiodic if the number of steps required to move between two states is not a multiple of some integer:

$$GCD\{n : p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}\} = 1.$$

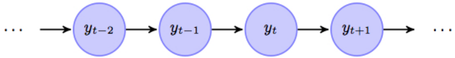
- Aperiodic Markov chains are NOT forced into cycles of fixed lengths between certain states.
- **Stationarity:** A Markov chain may reach stationarity when the probability of being in any particular state is independent of the initial state.
- The stationary distribution π satisfies

$$\pi = \pi \mathbf{P}.$$

- Irreducible and aperiodic Markov chains converge to stationarity.

Markov Chains - Stationarity

$$p(y_{t+1} \mid y_0, y_1, \dots, y_t) = p(y_{t+1} \mid y_t)$$



- **Irreducibility:** A Markov chain is irreducible if the chain can move from any state to any other in finite steps - there exists some $n \in \mathbb{N}$ such that
$$p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}.$$
- **Aperiodicity:** A Markov chain is aperiodic if the number of steps required to move between two states is not a multiple of some integer:
$$GCD\{n : p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}\} = 1.$$
- Aperiodic Markov chains are NOT forced into cycles of fixed lengths between certain states.
- **Stationarity:** A Markov chain may reach stationarity when the probability of being in any particular state is independent of the initial state.
- The stationary distribution π satisfies
$$\pi = \pi \mathbf{P}.$$
- Irreducible and aperiodic Markov chains converge to stationarity.

Markov Chains - Stationarity

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \longrightarrow y_{t-2} \longrightarrow y_{t-1} \longrightarrow y_t \longrightarrow y_{t+1} \longrightarrow \cdots$$

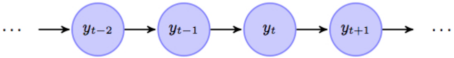
- **Irreducibility:** A Markov chain is irreducible if the chain can move from any state to any other in finite steps - there exists some $n \in \mathbb{N}$ such that
$$p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}.$$
- **Aperiodicity:** A Markov chain is aperiodic if the number of steps required to move between two states is not a multiple of some integer:
$$GCD\{n : p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}\} = 1.$$
- Aperiodic Markov chains are NOT forced into cycles of fixed lengths between certain states.
- **Stationarity:** A Markov chain may reach stationarity when the probability of being in any particular state is independent of the initial state.
- The stationary distribution π satisfies
$$\pi = \pi \mathbf{P}.$$
- Irreducible and aperiodic Markov chains converge to stationarity.

Markov Chains - Stationarity

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \to y_{t-2} \to y_{t-1} \to y_t \to y_{t+1} \to \cdots$$

- **Irreducibility:** A Markov chain is irreducible if the chain can move from any state to any other in finite steps - there exists some $n \in \mathbb{N}$ such that
$$p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}.$$
- **Aperiodicity:** A Markov chain is aperiodic if the number of steps required to move between two states is not a multiple of some integer:
$$GCD\{n : p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}\} = 1.$$
- Aperiodic Markov chains are NOT forced into cycles of fixed lengths between certain states.
- **Stationarity:** A Markov chain may reach stationarity when the probability of being in any particular state is independent of the initial state.
- The stationary distribution $\boldsymbol{\pi}$ satisfies
$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}.$$
- Irreducible and aperiodic Markov chains converge to stationarity.

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$$



- **Irreducibility:** A Markov chain is irreducible if the chain can move from any state to any other in finite steps - there exists some $n \in \mathbb{N}$ such that
$$p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}.$$
- **Aperiodicity:** A Markov chain is aperiodic if the number of steps required to move between two states is not a multiple of some integer:
$$GCD\{n : p^n(x, y) > 0 \quad \forall\, x, y \in \mathcal{Y}\} = 1.$$
- Aperiodic Markov chains are NOT forced into cycles of fixed lengths between certain states.
- **Stationarity:** A Markov chain may reach stationarity when the probability of being in any particular state is independent of the initial state.
- The stationary distribution $\boldsymbol{\pi}$ satisfies
$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}.$$
- Irreducible and aperiodic Markov chains converge to stationarity.

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \to y_{t-2} \to y_{t-1} \to y_t \to y_{t+1} \to \cdots$$

- $$\mathbf{P} = \begin{array}{c c} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0.4 & 0.4 & 0.2 \\ 0 & 0 & 0 & 0.2 & 0.8 \end{bmatrix} \end{array}, \qquad \begin{array}{l} 1 \leftrightarrow 2, \\ 3 \leftrightarrow 4 \leftrightarrow 5. \end{array}$$

- $$\mathbf{P} = \begin{array}{c c} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix} \end{array}, \qquad \pi\mathbf{P} = \pi \quad \Rightarrow \quad \pi = (0.5, 0.5).$$

- $$\mathbf{P} = \begin{array}{c c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array} \rightarrow \text{irreducible, periodic Markov chain with period 4.}$$

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \to y_{t-2} \to y_{t-1} \to y_t \to y_{t+1} \to \cdots$$

- $$\mathbf{P} = \begin{array}{c c} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0.4 & 0.4 & 0.2 \\ 0 & 0 & 0 & 0.2 & 0.8 \end{bmatrix} \end{array}, \qquad \begin{array}{l} 1 \leftrightarrow 2, \\ 3 \leftrightarrow 4 \leftrightarrow 5. \end{array}$$

- $$\mathbf{P} = \begin{array}{c c} & \begin{array}{cc} 1 & 2 \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} & \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix} \end{array}, \qquad \boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi} \quad \Rightarrow \quad \boldsymbol{\pi} = (0.5, 0.5).$$

- $$\mathbf{P} = \begin{array}{c c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array} \quad \rightarrow \text{ irreducible, periodic Markov chain with period 4.}$$

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \to y_{t-2} \to y_{t-1} \to y_t \to y_{t+1} \to \cdots$$

- $$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0.4 & 0.4 & 0.2 \\ 0 & 0 & 0 & 0.2 & 0.8 \end{bmatrix} \end{array}, \qquad \begin{array}{l} 1 \leftrightarrow 2, \\ 3 \leftrightarrow 4 \leftrightarrow 5. \end{array}$$

- $$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix} \end{array}, \qquad \boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi} \quad \Rightarrow \quad \boldsymbol{\pi} = (0.5, 0.5).$$

- $$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array} \rightarrow$$ irreducible, periodic Markov chain with period 4.

Markov Chains - Reversibility

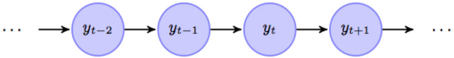
$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \longrightarrow y_{t-2} \longrightarrow y_{t-1} \longrightarrow y_t \longrightarrow y_{t+1} \longrightarrow \cdots$$

- **Detailed Balance Equation:** $\pi^\star(x)p(x,y) = \pi^\star(y)p(y,x) \;\; \forall\, (x,y) \in \mathcal{Y}^2$.
- **Reversibility:** A Markov chain is reversible if the detailed balanced equation holds.
- Detailed balance equation implies $\boldsymbol{\pi}^\star = \boldsymbol{\pi}^\star \mathbf{P}$.

  $(\boldsymbol{\pi}^\star \mathbf{P})_y = \sum_x \pi^\star(x)p(x,y) = \sum_x \pi^\star(y)p(y,x) = \pi^\star(y)\sum_x p(y,x) = \pi^\star(y)$.
- This implies $\boldsymbol{\pi}^\star = \boldsymbol{\pi}$, the stationary distribution of $\mathbf{P}$.

Markov Chains - Reversibility

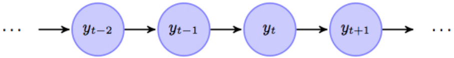
$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$



- **Detailed Balance Equation:** $\pi^\star(x)p(x,y) = \pi^\star(y)p(y,x) \;\; \forall\, (x,y) \in \mathcal{Y}^2$.
- **Reversibility:** A Markov chain is reversible if the detailed balanced equation holds.
- Detailed balance equation implies $\boldsymbol{\pi}^\star = \boldsymbol{\pi}^\star \mathbf{P}$.

  $(\boldsymbol{\pi}^\star \mathbf{P})_y = \sum_x \pi^\star(x)p(x,y) = \sum_x \pi^\star(y)p(y,x) = \pi^\star(y)\sum_x p(y,x) = \pi^\star(y)$.
- This implies $\boldsymbol{\pi}^\star = \boldsymbol{\pi}$, the stationary distribution of $\mathbf{P}$.

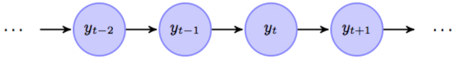
$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$



- **Detailed Balance Equation:** $\pi^\star(x)p(x,y) = \pi^\star(y)p(y,x) \;\; \forall\, (x,y) \in \mathcal{Y}^2$.
- **Reversibility:** A Markov chain is reversible if the detailed balanced equation holds.
- Detailed balance equation implies $\boldsymbol{\pi}^\star = \boldsymbol{\pi}^\star \mathbf{P}$.

  $(\boldsymbol{\pi}^\star \mathbf{P})_y = \sum_x \pi^\star(x)p(x,y) = \sum_x \pi^\star(y)p(y,x) = \pi^\star(y) \sum_x p(y,x) = \pi^\star(y)$.
- This implies $\boldsymbol{\pi}^\star = \boldsymbol{\pi}$, the stationary distribution of $\mathbf{P}$.

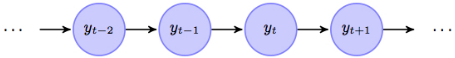
$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$$



- **Detailed Balance Equation:** $\pi^\star(x)p(x,y) = \pi^\star(y)p(y,x) \;\; \forall \, (x,y) \in \mathcal{Y}^2$.
- **Reversibility:** A Markov chain is reversible if the detailed balanced equation holds.
- Detailed balance equation implies $\boldsymbol{\pi}^\star = \boldsymbol{\pi}^\star \mathbf{P}$.

  $(\boldsymbol{\pi}^\star \mathbf{P})_y = \sum_x \pi^\star(x)p(x,y) = \sum_x \pi^\star(y)p(y,x) = \pi^\star(y)\sum_x p(y,x) = \pi^\star(y)$.
- This implies $\boldsymbol{\pi}^\star = \boldsymbol{\pi}$, the stationary distribution of $\mathbf{P}$.

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$$



- **Detailed Balance Equation:** $\pi^\star(x)p(x,y) = \pi^\star(y)p(y,x) \;\; \forall\, (x,y) \in \mathcal{Y}^2$.
- **Reversibility:** A Markov chain is reversible if the detailed balanced equation holds.
- Detailed balance equation implies $\boldsymbol{\pi}^\star = \boldsymbol{\pi}^\star \mathbf{P}$.

  $(\boldsymbol{\pi}^\star \mathbf{P})_y = \sum_x \pi^\star(x)p(x,y) = \sum_x \pi^\star(y)p(y,x) = \pi^\star(y)\sum_x p(y,x) = \pi^\star(y)$.
- This implies $\boldsymbol{\pi}^\star = \boldsymbol{\pi}$, the stationary distribution of $\mathbf{P}$.

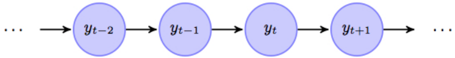
Markov Chains - Continuous Case

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t) \qquad \cdots \to y_{t-2} \to y_{t-1} \to y_t \to y_{t+1} \to \cdots$$

- **Transition Probability Kernel:** $\int p(x, y)dy = 1.$
- **Chapman-Kolmogorov Equation:** $\pi_t(y) = \int \pi_{t-1}(x)p(x, y)dx.$
- **Detailed Balance Equation:** $\pi(x)p(x, y) = \pi(y)p(y, x) \;\; \forall \; (x, y) \in \mathcal{Y}^2.$
- **Stationary Distribution:** $\pi(y) = \int \pi(x)p(x, y)dx.$

Markov Chains - Continuous Case

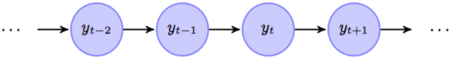
$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$$



- **Transition Probability Kernel:** $\int p(x, y)dy = 1$.
- **Chapman-Kolmogorov Equation:** $\pi_t(y) = \int \pi_{t-1}(x)p(x, y)dx$.
- **Detailed Balance Equation:** $\pi(x)p(x, y) = \pi(y)p(y, x) \;\; \forall\, (x, y) \in \mathcal{Y}^2$.
- **Stationary Distribution:** $\pi(y) = \int \pi(x)p(x, y)dx$.

Markov Chains - Continuous Case

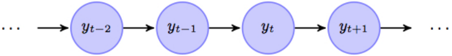
$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$



- **Transition Probability Kernel:** $\int p(x, y)dy = 1$.
- **Chapman-Kolmogorov Equation:** $\pi_t(y) = \int \pi_{t-1}(x)p(x, y)dx$.
- **Detailed Balance Equation:** $\pi(x)p(x, y) = \pi(y)p(y, x) \;\; \forall \; (x, y) \in \mathcal{Y}^2$.
- **Stationary Distribution:** $\pi(y) = \int \pi(x)p(x, y)dx$.

Markov Chains - Continuous Case

$$p(y_{t+1} \mid y_0, y_1, \ldots, y_t) = p(y_{t+1} \mid y_t)$$



- **Transition Probability Kernel:** $\int p(x,y)dy = 1.$
- **Chapman-Kolmogorov Equation:** $\pi_t(y) = \int \pi_{t-1}(x)p(x,y)dx.$
- **Detailed Balance Equation:** $\pi(x)p(x,y) = \pi(y)p(y,x) \;\; \forall \, (x,y) \in \mathcal{Y}^2.$
- **Stationary Distribution:** $\pi(y) = \int \pi(x)p(x,y)dx.$

Metropolis-Hastings Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)$$

$$\cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- Goal is to draw random samples from a (generic) distribution $p(\boldsymbol{\theta})$ possibly known only up to its normalizing constant, e.g., a complex joint posterior $p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})$.
- **Iterative Algorithm (Metropolis Sampler):**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  (a) Generate a candidate $\boldsymbol{\theta}^{\star}$ using a 'proposal distribution' $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^{\star})$ satisfying $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^{\star}) = q(\boldsymbol{\theta}^{\star} \to \boldsymbol{\theta}^{(t-1)})$.
  (b) Compute $\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^{\star})}{p(\boldsymbol{\theta}^{(t-1)})}\right\}$.
  (c) With probability α, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{\star}$. With probability $(1-\alpha)$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
- **Iterative Algorithm (Metropolis-Hastings Sampler):**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  (a) Generate a candidate $\boldsymbol{\theta}^{\star}$ using a 'proposal distribution' $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^{\star})$.
  (b) Compute $\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^{\star})q(\boldsymbol{\theta}^{\star} \to \boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^{\star})}\right\}$.
  (c) With probability α, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{\star}$. With probability $(1-\alpha)$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
- **Step (c):** Draw $r \sim \text{Unif}(0,1)$. If $r \leq \alpha$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{\star}$, otherwise, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

Metropolis-Hastings Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- Goal is to draw random samples from a (generic) distribution $p(\boldsymbol{\theta})$ possibly known only up to its normalizing constant, e.g., a complex joint posterior $p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})$.

- **Iterative Algorithm (Metropolis Sampler):**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  (a) Generate a candidate $\boldsymbol{\theta}^{\star}$ using a 'proposal distribution' $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^{\star})$ satisfying $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^{\star}) = q(\boldsymbol{\theta}^{\star} \to \boldsymbol{\theta}^{(t-1)})$.
  (b) Compute $\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^{\star})}{p(\boldsymbol{\theta}^{(t-1)})}\right\}$.
  (c) With probability α, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{\star}$. With probability $(1-\alpha)$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

- **Iterative Algorithm (Metropolis-Hastings Sampler):**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  (a) Generate a candidate $\boldsymbol{\theta}^{\star}$ using a 'proposal distribution' $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^{\star})$.
  (b) Compute $\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^{\star})q(\boldsymbol{\theta}^{\star} \to \boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^{\star})}\right\}$.
  (c) With probability α, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{\star}$. With probability $(1-\alpha)$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

- **Step (c):** Draw $r \sim \text{Unif}(0,1)$. If $r \leq \alpha$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{\star}$, otherwise, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

Metropolis-Hastings Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- Goal is to draw random samples from a (generic) distribution $p(\boldsymbol{\theta})$ possibly known only up to its normalizing constant, e.g., a complex joint posterior $p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})$.
- **Iterative Algorithm (Metropolis Sampler):**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  (a) Generate a candidate $\boldsymbol{\theta}^\star$ using a 'proposal distribution' $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star)$ satisfying $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star) = q(\boldsymbol{\theta}^\star \to \boldsymbol{\theta}^{(t-1)})$.
  (b) Compute $\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^\star)}{p(\boldsymbol{\theta}^{(t-1)})}\right\}$.
  (c) With probability α, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^\star$. With probability $(1-\alpha)$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
- **Iterative Algorithm (Metropolis-Hastings Sampler):**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  (a) Generate a candidate $\boldsymbol{\theta}^\star$ using a 'proposal distribution' $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star)$.
  (b) Compute $\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^\star)q(\boldsymbol{\theta}^\star \to \boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star)}\right\}$.
  (c) With probability α, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^\star$. With probability $(1-\alpha)$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
- **Step (c):** Draw $r \sim \text{Unif}(0,1)$. If $r \le \alpha$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^\star$, otherwise, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

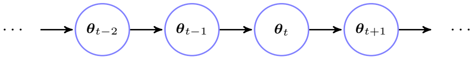
Metropolis-Hastings Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \to \boldsymbol{\theta}_{t-2} \to \boldsymbol{\theta}_{t-1} \to \boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1} \to \cdots$$

- Goal is to draw random samples from a (generic) distribution $p(\boldsymbol{\theta})$ possibly known only up to its normalizing constant, e.g., a complex joint posterior $p(\boldsymbol{\theta} \mid \mathbf{y}_{1:n})$.
- **Iterative Algorithm (Metropolis Sampler):**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  (a) Generate a candidate $\boldsymbol{\theta}^\star$ using a 'proposal distribution' $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star)$ satisfying $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star) = q(\boldsymbol{\theta}^\star \to \boldsymbol{\theta}^{(t-1)})$.
  (b) Compute $\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^\star)}{p(\boldsymbol{\theta}^{(t-1)})}\right\}$.
  (c) With probability α, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^\star$. With probability $(1-\alpha)$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
- **Iterative Algorithm (Metropolis-Hastings Sampler):**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  (a) Generate a candidate $\boldsymbol{\theta}^\star$ using a 'proposal distribution' $q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star)$.
  (b) Compute $\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^\star) q(\boldsymbol{\theta}^\star \to \boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)}) q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star)}\right\}$.
  (c) With probability α, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^\star$. With probability $(1-\alpha)$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
- **Step (c):** Draw $r \sim \mathrm{Unif}(0,1)$. If $r \leq \alpha$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^\star$, otherwise, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

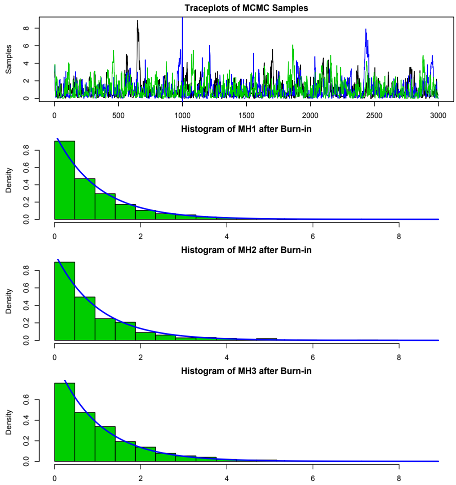
Metropolis-Hastings Sampler - Sampling from an Exponential

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)$$



- Target: $p(\theta) = \exp(-\theta)$
- Proposal: $q(\theta^\star|\theta^{(t-1)}) = \text{Normal}(\theta^\star \mid \theta^{(t-1)}, \sigma_\theta^2)$

- Start with some $\theta^{(0)}$.
- Propose $\theta^\star$ according to $q(\theta^\star|\theta^{(t-1)})$.
- Compute $\alpha = \min\left\{1, \frac{p(\theta^\star)}{p(\theta^{(t-1)})}\right\}$.
- Draw $r = \text{Unif}(0, 1)$.
- If $r \leq \alpha$, set $\theta^{(t)} = \theta^\star$, otherwise, set $\theta^{(t)} = \theta^{(t-1)}$.



Metropolis-Hastings Sampler as a Markov Chain

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \to \boldsymbol{\theta}_{t-2} \to \boldsymbol{\theta}_{t-1} \to \boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1} \to \cdots$$

- The transition probability kernel of the MH sampler is

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

- For all $(x, y) \in \mathcal{Y}^2$, either $\alpha(x, y) = 1$ or $\alpha(y, x) = 1$. If $\alpha(x, y) = 1$, then

$$\begin{aligned} \alpha(y, x) &= \frac{p(x)q(x,y)}{p(y)q(y,x)} = \frac{p(x)q(x,y)\alpha(x,y)}{p(y)q(y,x)} \\ &\Leftrightarrow \alpha(y, x)p(y)q(y, x) = p(x)q(x, y)\alpha(x, y) \\ &\Leftrightarrow p(y)p(y, x) = p(x)p(x, y). \end{aligned}$$

- MH sampler constructs a Markov chain with stationary distribution $p(y)$.
- Irrespective of initial values, when run long enough, MH sampler eventually draws samples from the stationary distribution $p(y)$.

Metropolis-Hastings Sampler as a Markov Chain

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_{t} \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- The transition probability kernel of the MH sampler is

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

- For all $(x, y) \in \mathcal{Y}^2$, either $\alpha(x, y) = 1$ or $\alpha(y, x) = 1$. If $\alpha(x, y) = 1$, then

$$\begin{aligned}
\alpha(y, x) &= \frac{p(x)q(x,y)}{p(y)q(y,x)} = \frac{p(x)q(x,y)\alpha(x,y)}{p(y)q(y,x)} \\
&\Leftrightarrow \alpha(y, x)p(y)q(y, x) = p(x)q(x, y)\alpha(x, y) \\
&\Leftrightarrow p(y)p(y, x) = p(x)p(x, y).
\end{aligned}$$

- MH sampler constructs a Markov chain with stationary distribution $p(y)$.
- Irrespective of initial values, when run long enough, MH sampler eventually draws samples from the stationary distribution $p(y)$.

Metropolis-Hastings Sampler as a Markov Chain

$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)$ $\cdots \to \boldsymbol{\theta}_{t-2} \to \boldsymbol{\theta}_{t-1} \to \boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1} \to \cdots$

- The transition probability kernel of the MH sampler is

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

- For all $(x, y) \in \mathcal{Y}^2$, either $\alpha(x, y) = 1$ or $\alpha(y, x) = 1$. If $\alpha(x, y) = 1$, then

$$\begin{aligned} \alpha(y, x) &= \frac{p(x)q(x,y)}{p(y)q(y,x)} = \frac{p(x)q(x,y)\alpha(x,y)}{p(y)q(y,x)} \\ &\Leftrightarrow \alpha(y, x)p(y)q(y, x) = p(x)q(x, y)\alpha(x, y) \\ &\Leftrightarrow p(y)p(y, x) = p(x)p(x, y). \end{aligned}$$

- MH sampler constructs a Markov chain with stationary distribution $p(y)$.
- Irrespective of initial values, when run long enough, MH sampler eventually draws samples from the stationary distribution $p(y)$.

Metropolis-Hastings Sampler as a Markov Chain

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

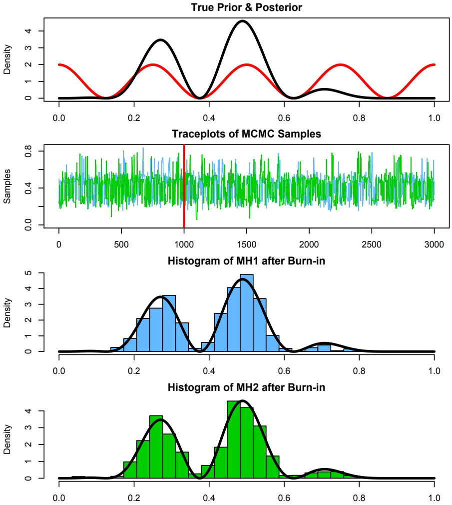
- The transition probability kernel of the MH sampler is
$$p(x, y) = q(x, y)\alpha(x, y) = q(x, y) \min\left\{\frac{p(y)q(y,x)}{p(x)q(x,y)}, 1\right\}.$$
- For all $(x, y) \in \mathcal{Y}^2$, either $\alpha(x, y) = 1$ or $\alpha(y, x) = 1$. If $\alpha(x, y) = 1$, then
$$\begin{aligned} \alpha(y, x) &= \frac{p(x)q(x,y)}{p(y)q(y,x)} = \frac{p(x)q(x,y)\alpha(x,y)}{p(y)q(y,x)} \\ &\Leftrightarrow \alpha(y, x)p(y)q(y, x) = p(x)q(x, y)\alpha(x, y) \\ &\Leftrightarrow p(y)p(y, x) = p(x)p(x, y). \end{aligned}$$
- MH sampler constructs a Markov chain with stationary distribution $p(y)$.
- Irrespective of initial values, when run long enough, MH sampler eventually draws samples from the stationary distribution $p(y)$.

Metropolis-Hastings Sampler - Sampling from a Complex Posterior

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- Likelihood: $p(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$
- Prior: $p(\theta) \propto \cos^2(4\pi\theta)$
- Target: $p(\theta \mid \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta)\ \theta^s(1-\theta)^{n-s}$
- Proposal: $q(\theta^\star|\theta^{(t-1)}) = \text{Normal}(\theta^\star \mid \theta^{(t-1)}, \sigma_\theta^2)$

- Start with some $\theta^{(0)}$.
- Propose $\theta^\star$ according to $q(\theta^\star|\theta^{(t-1)})$.
- Compute $\alpha = \min\left\{1, \frac{p(\theta^\star|\mathbf{y}_{1:n})}{p(\theta^{(t-1)}|\mathbf{y}_{1:n})}\right\}$.
- Draw $r = \text{Unif}(0,1)$.
- If $r \leq \alpha$, set $\theta^{(t)} = \theta^\star$, otherwise, set $\theta^{(t)} = \theta^{(t-1)}$.

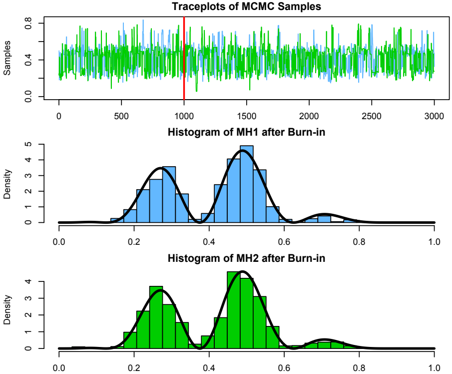


Metropolis-Hastings Sampler - Mixing

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \quad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- **Good Mixing:** The chain explores the whole parameter space well.
- **Poor Mixing 1:** The chain stays in small regions of the parameter space.
- **Poor Mixing 2:** The chain makes big jumps with little acceptance.

- Likelihood: $p(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$
- Prior: $p(\theta) \propto \cos^2(4\pi\theta)$
- Target: $p(\theta \mid \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta)\ \theta^s(1-\theta)^{n-s}$
- Proposal: $q(\theta^\star|\theta^{(t-1)}) = \text{Normal}(\theta^\star \mid \theta^{(t-1)}, \mathbf{0.25^2})$
- Starting points: MH1: $\theta^{(0)} = 0.50$, MH2: $\theta^{(0)} = 0.25$

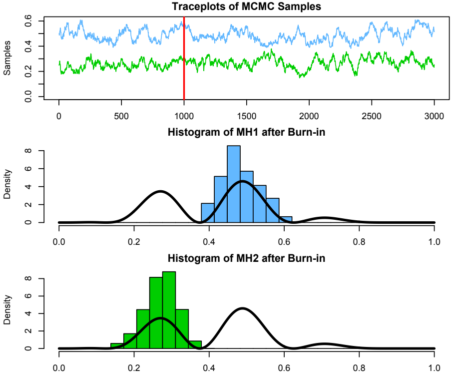


Metropolis-Hastings Sampler - Mixing

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \to \boldsymbol{\theta}_{t-2} \to \boldsymbol{\theta}_{t-1} \to \boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1} \to \cdots$$

- **Good Mixing:** The chain explores the whole parameter space well.
- **Poor Mixing 1:** The chain stays in small regions of the parameter space.
- **Poor Mixing 2:** The chain makes big jumps with little acceptance.

- Likelihood: $p(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$
- Prior: $p(\theta) \propto \cos^2(4\pi\theta)$
- Target: $p(\theta \mid \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta)\ \theta^s(1-\theta)^{n-s}$
- Proposal: $q(\theta^\star|\theta^{(t-1)}) = \text{Normal}(\theta^\star \mid \theta^{(t-1)}, \mathbf{0.01^2})$
- Starting points: MH1: $\theta^{(0)} = 0.50$, MH2: $\theta^{(0)} = 0.25$

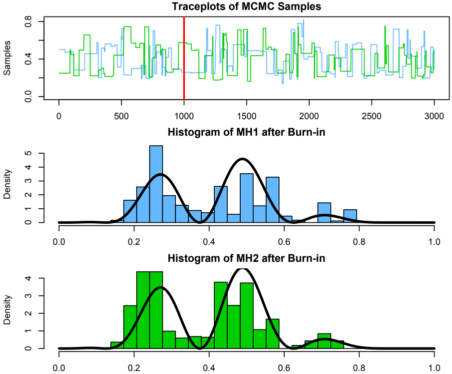


Metropolis-Hastings Sampler - Mixing

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \quad \cdots \rightarrow \boldsymbol{\theta}_{t-2} \rightarrow \boldsymbol{\theta}_{t-1} \rightarrow \boldsymbol{\theta}_t \rightarrow \boldsymbol{\theta}_{t+1} \rightarrow \cdots$$

- **Good Mixing:** The chain explores the whole parameter space well.
- **Poor Mixing 1:** The chain stays in small regions of the parameter space.
- **Poor Mixing 2:** The chain makes big jumps with little acceptance.

- Likelihood: $p(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$
- Prior: $p(\theta) \propto \cos^2(4\pi\theta)$
- Target: $p(\theta \mid \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta)\ \theta^s(1-\theta)^{n-s}$
- Proposal: $q(\theta^\star|\theta^{(t-1)}) = \text{Normal}(\theta^\star \mid \theta^{(t-1)}, \mathbf{4^2})$
- Starting points: MH1: $\theta^{(0)} = 0.50$, MH2: $\theta^{(0)} = 0.25$



Metropolis-Hastings Sampler - Simulated Annealing

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \to \boldsymbol{\theta}_{t-2} \to \boldsymbol{\theta}_{t-1} \to \boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1} \to \cdots$$

Metropolis-Hastings Sampler:

- Start with some $\boldsymbol{\theta}^{(0)}$.
- Propose $\boldsymbol{\theta}^\star$ according to $q(\boldsymbol{\theta}^\star | \boldsymbol{\theta}^{(t-1)})$.
- Compute $\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^\star) q(\boldsymbol{\theta}^\star \to \boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)}) q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star)}\right\}$
- Draw $r = \mathrm{Unif}(0, 1)$.
- If $r \leq \alpha$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^\star$, otherwise, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

Metropolis-Hastings Sampler with Simulated Annealing:

- Modify the acceptance probability as $\alpha = \min\left\{1, \left\{\frac{p(\boldsymbol{\theta}^\star) q(\boldsymbol{\theta}^\star \to \boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)}) q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star)}\right\}^{1/T(t)}\right\}$
  - Start with temperature T_0 and cool down to a final temperature T_f over n iterations: $T(t) = \max\left\{T_0 \left(\frac{T_f}{T_0}\right)^{t/n}, T_f\right\}$.
  - Start with temperature T_0 and cool down to original MH over n iterations: $T(t) = \max\left\{T_0^{1-t/n}, 1\right\}$.

Metropolis-Hastings Sampler - Simulated Annealing

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \quad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_{t} \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \quad \cdots$$

Metropolis-Hastings Sampler:

- Start with some $\boldsymbol{\theta}^{(0)}$.
- Propose $\boldsymbol{\theta}^\star$ according to $q(\boldsymbol{\theta}^\star | \boldsymbol{\theta}^{(t-1)})$.
- Compute $\alpha = \min \left\{ 1, \frac{p(\boldsymbol{\theta}^\star) q(\boldsymbol{\theta}^\star \to \boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)}) q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star)} \right\}$
- Draw $r = \mathrm{Unif}(0, 1)$.
- If $r \leq \alpha$, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^\star$, otherwise, set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

Metropolis-Hastings Sampler with Simulated Annealing:

- Modify the acceptance probability as $\alpha = \min \left\{ 1, \left\{ \frac{p(\boldsymbol{\theta}^\star) q(\boldsymbol{\theta}^\star \to \boldsymbol{\theta}^{(t-1)})}{p(\boldsymbol{\theta}^{(t-1)}) q(\boldsymbol{\theta}^{(t-1)} \to \boldsymbol{\theta}^\star)} \right\}^{1/T(t)} \right\}$
    - Start with temperature T_0 and cool down to a final temperature T_f over n iterations:
      $T(t) = \max \left\{ T_0 \left(\frac{T_f}{T_0} \right)^{t/n}, T_f \right\}$.
    - Start with temperature T_0 and cool down to original MH over n iterations:
      $T(t) = \max \left\{ T_0^{1-t/n}, 1 \right\}$.

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- **Autocorrelation:** The k^{th} order autocorrelation based on n samples after burn-in:

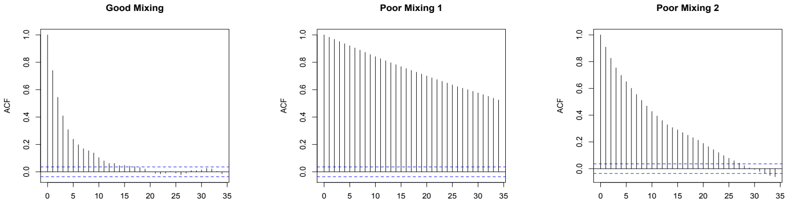
$$\rho_k = \frac{\mathrm{cov}(\theta^{(t)}, \theta^{(t+k)})}{\mathrm{var}(\theta^{(t)})} \widehat{=} \frac{\sum_{t=1}^{n-k}(\theta^{(t)} - \overline{\theta})(\theta^{(t+k)} - \overline{\theta})}{\sum_{t=1}^{n-k}(\theta^{(t)} - \overline{\theta})^2} = \widehat{\rho}_k, \qquad \overline{\theta} = \frac{1}{n}\sum_{t=1}^{n} \theta^{(t)}.$$

Convergence Diagnostics

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_{t} \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- **Autocorrelation:** The k^{th} order autocorrelation based on n samples after burn-in:

$$\rho_k = \frac{\mathrm{cov}(\theta^{(t)}, \theta^{(t+k)})}{\mathrm{var}(\theta^{(t)})} \;\widehat{=}\; \frac{\sum_{t=1}^{n-k}(\theta^{(t)} - \overline{\theta})(\theta^{(t+k)} - \overline{\theta})}{\sum_{t=1}^{n-k}(\theta^{(t)} - \overline{\theta})^2} = \widehat{\rho}_k, \qquad \overline{\theta} = \frac{1}{n}\sum_{t=1}^{n} \theta^{(t)}.$$



$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- **Autocorrelation:** The k^{th} order autocorrelation based on n samples after burn-in:

$$\rho_k = \frac{\mathrm{cov}(\theta^{(t)}, \theta^{(t+k)})}{\mathrm{var}(\theta^{(t)})} \mathrel{\widehat{=}} \frac{\sum_{t=1}^{n-k} (\theta^{(t)} - \overline{\theta})(\theta^{(t+k)} - \overline{\theta})}{\sum_{t=1}^{n-k} (\theta^{(t)} - \overline{\theta})^2} = \widehat{\rho}_k, \qquad \overline{\theta} = \frac{1}{n} \sum_{t=1}^{n} \theta^{(t)}.$$

- **Sample Size Inflation:** For an AR(1) process $\theta^{(t)} = \mu + \alpha(\theta^{(t-1)} - \mu) + \epsilon_t$ with $\epsilon_t \overset{iid}{\sim} \mathrm{Normal}(0, \sigma^2)$, we have

$$\rho_k = \alpha^k, \qquad \mathbb{E}(\overline{\theta}_T) = \mu, \qquad SE(\overline{\theta}_T) \approx \frac{\sigma}{\sqrt{T}} \frac{1}{(1-\rho)}.$$

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- **Autocorrelation:** The k^{th} order autocorrelation based on n samples after burn-in:

$$\rho_k = \frac{\text{cov}(\theta^{(t)}, \theta^{(t+k)})}{\text{var}(\theta^{(t)})} \mathrel{\widehat{=}} \frac{\sum_{t=1}^{n-k}(\theta^{(t)} - \overline{\theta})(\theta^{(t+k)} - \overline{\theta})}{\sum_{t=1}^{n-k}(\theta^{(t)} - \overline{\theta})^2} = \widehat{\rho}_k, \qquad \overline{\theta} = \frac{1}{n}\sum_{t=1}^{n} \theta^{(t)}.$$

- **Sample Size Inflation:** For an AR(1) process $\theta^{(t)} = \mu + \alpha(\theta^{(t-1)} - \mu) + \epsilon_t$ with $\epsilon_t \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$, we have

$$\rho_k = \alpha^k, \qquad \mathbb{E}(\overline{\theta}_T) = \mu, \qquad SE(\overline{\theta}_T) \approx \frac{\sigma}{\sqrt{T}} \frac{1}{(1-\rho)}.$$

  Therefore, $\overline{\theta}$ is unbiased for μ but with inflated variance.

- **Thinning:** Thin the samples after burn-in, e.g., take every k^{th} value or

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_{t} \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

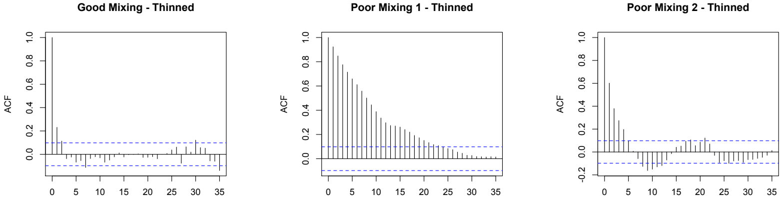
- **Autocorrelation:** The k^{th} order autocorrelation based on n samples after burn-in:

$$\rho_k = \frac{\mathrm{cov}(\theta^{(t)}, \theta^{(t+k)})}{\mathrm{var}(\theta^{(t)})} \;\widehat{=}\; \frac{\sum_{t=1}^{n-k}(\theta^{(t)} - \bar{\theta})(\theta^{(t+k)} - \bar{\theta})}{\sum_{t=1}^{n-k}(\theta^{(t)} - \bar{\theta})^2} = \widehat{\rho}_k, \qquad \bar{\theta} = \frac{1}{n}\sum_{t=1}^{n} \theta^{(t)}.$$

- **Sample Size Inflation:** For an AR(1) process $\theta^{(t)} = \mu + \alpha(\theta^{(t-1)} - \mu) + \epsilon_t$ with $\epsilon_t \overset{iid}{\sim} \mathrm{Normal}(0, \sigma^2)$, we have

$$\rho_k = \alpha^k, \qquad \mathbb{E}(\bar{\theta}_T) = \mu, \qquad SE(\bar{\theta}_T) \approx \frac{\sigma}{\sqrt{T}} \frac{1}{(1-\rho)}.$$

- **Thinning:** Thin the samples after burn-in, e.g. take every 5^{th} value etc.



Gibbs Sampler

$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)$ $\quad \cdots \rightarrow \boldsymbol{\theta}_{t-2} \rightarrow \boldsymbol{\theta}_{t-1} \rightarrow \boldsymbol{\theta}_t \rightarrow \boldsymbol{\theta}_{t+1} \rightarrow \cdots$

- Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ and $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)^{\mathrm{T}}$ for each j.
- Sampling from the p-dimensional joint (posterior) distribution is often difficult.
- Computing smaller dimensional marginal (posterior) distributions is often difficult.

$$p(\theta_j) = \int p(\theta_j, \boldsymbol{\theta}_{-j}) d\boldsymbol{\theta}_{-j}$$

- Computing smaller dimensional conditional (posterior) distributions is often easy.

$$p(\theta_j \mid \boldsymbol{\theta}_{-j}) \propto p(\theta_j, \boldsymbol{\theta}_{-j}) \quad \text{with } \boldsymbol{\theta}_{-j} \text{ held constant}$$

- Special type of MH sampler when the proposal is never rejected ($\alpha = 1$).
- Samples only from smaller dimensional conditional (posterior) distributions.
- **Iterative Algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  At each iteration t, for each $j = 1, \ldots, p$, sample $\theta_j^{(t)}$ from

$$p(\theta_j^{(t)} \mid \theta_1 = \theta_1^{(t)}, \cdots, \theta_{j-1} = \theta_{j-1}^{(t)}, \theta_{j+1} = \theta_{j+1}^{(t-1)}, \ldots, \theta_p = \theta_p^{(t-1)}).$$

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ and $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)^{\mathrm{T}}$ for each j.
- Sampling from the p-dimensional joint (posterior) distribution is often difficult.
- Computing smaller dimensional marginal (posterior) distributions is often difficult.

$$p(\theta_j) = \int p(\theta_j, \boldsymbol{\theta}_{-j}) d\boldsymbol{\theta}_{-j}$$

- Computing smaller dimensional conditional (posterior) distributions is often easy.

$$p(\theta_j \mid \boldsymbol{\theta}_{-j}) \propto p(\theta_j, \boldsymbol{\theta}_{-j}) \quad \text{with } \boldsymbol{\theta}_{-j} \text{ held constant}$$

- Special type of MH sampler when the proposal is never rejected ($\alpha = 1$).
- Samples only from smaller dimensional conditional (posterior) distributions.
- **Iterative Algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  At each iteration t, for each $j = 1, \ldots, p$, sample $\theta_j^{(t)}$ from

$$p(\theta_j^{(t)} \mid \theta_1 = \theta_1^{(t)}, \cdots, \theta_{j-1} = \theta_{j-1}^{(t)}, \theta_{j+1} = \theta_{j+1}^{(t-1)}, \ldots, \theta_p = \theta_p^{(t-1)}).$$

Gibbs Sampler

$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)$ $\quad \cdots \to \boldsymbol{\theta}_{t-2} \to \boldsymbol{\theta}_{t-1} \to \boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1} \to \cdots$

- Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ and $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)^{\mathrm{T}}$ for each j.
- Sampling from the p-dimensional joint (posterior) distribution is often difficult.
- Computing smaller dimensional marginal (posterior) distributions is often difficult.

$$p(\theta_j) = \int p(\theta_j, \boldsymbol{\theta}_{-j}) d\boldsymbol{\theta}_{-j}$$

- Computing smaller dimensional conditional (posterior) distributions is often easy.

$$p(\theta_j \mid \boldsymbol{\theta}_{-j}) \propto p(\theta_j, \boldsymbol{\theta}_{-j}) \quad \text{with } \boldsymbol{\theta}_{-j} \text{ held constant}$$

- Special type of MH sampler when the proposal is never rejected ($\alpha = 1$).
- Samples only from smaller dimensional conditional (posterior) distributions.
- **Iterative Algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  At each iteration t, for each $j = 1, \ldots, p$, sample $\theta_j^{(t)}$ from

$$p(\theta_j^{(t)} \mid \theta_1 = \theta_1^{(t)}, \cdots, \theta_{j-1} = \theta_{j-1}^{(t)}, \theta_{j+1} = \theta_{j+1}^{(t-1)}, \ldots, \theta_p = \theta_p^{(t-1)}).$$

Gibbs Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \to \boldsymbol{\theta}_{t-2} \to \boldsymbol{\theta}_{t-1} \to \boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1} \to \cdots$$

- Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ and $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)^{\mathrm{T}}$ for each j.
- Sampling from the p-dimensional joint (posterior) distribution is often difficult.
- Computing smaller dimensional marginal (posterior) distributions is often difficult.

$$p(\theta_j) = \int p(\theta_j, \boldsymbol{\theta}_{-j}) d\boldsymbol{\theta}_{-j}$$

- Computing smaller dimensional conditional (posterior) distributions is often easy.

$$p(\theta_j \mid \boldsymbol{\theta}_{-j}) \propto p(\theta_j, \boldsymbol{\theta}_{-j}) \quad \text{with } \boldsymbol{\theta}_{-j} \text{ held constant}$$

- Special type of MH sampler when the proposal is never rejected ($\alpha = 1$).
- Samples only from smaller dimensional conditional (posterior) distributions.
- **Iterative Algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  At each iteration t, for each $j = 1, \ldots, p$, sample $\theta_j^{(t)}$ from

$$p(\theta_j^{(t)} \mid \theta_1 = \theta_1^{(t)}, \cdots, \theta_{j-1} = \theta_{j-1}^{(t)}, \theta_{j+1} = \theta_{j+1}^{(t-1)}, \ldots, \theta_p = \theta_p^{(t-1)}).$$

Gibbs Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \to \boldsymbol{\theta}_{t-2} \to \boldsymbol{\theta}_{t-1} \to \boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1} \to \cdots$$

- Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ and $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)^{\mathrm{T}}$ for each j.
- Sampling from the p-dimensional joint (posterior) distribution is often difficult.
- Computing smaller dimensional marginal (posterior) distributions is often difficult.

$$p(\theta_j) = \int p(\theta_j, \boldsymbol{\theta}_{-j}) d\boldsymbol{\theta}_{-j}$$

- Computing smaller dimensional conditional (posterior) distributions is often easy.

$$p(\theta_j \mid \boldsymbol{\theta}_{-j}) \propto p(\theta_j, \boldsymbol{\theta}_{-j}) \quad \text{with } \boldsymbol{\theta}_{-j} \text{ held constant}$$

Gibbs Sampler:

- Special type of MH sampler when the proposal is never rejected ($\alpha = 1$).
- Samples only from smaller dimensional conditional (posterior) distributions.
- **Iterative Algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  At each iteration t, for each $j = 1, \ldots, p$, sample $\theta_j^{(t)}$ from

$$p(\theta_j^{(t)} \mid \theta_1 = \theta_1^{(t)}, \cdots, \theta_{j-1} = \theta_{j-1}^{(t)}, \theta_{j+1} = \theta_{j+1}^{(t-1)}, \ldots, \theta_p = \theta_p^{(t-1)}).$$

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \to \boldsymbol{\theta}_{t-2} \to \boldsymbol{\theta}_{t-1} \to \boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1} \to \cdots$$

- Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ and $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)^{\mathrm{T}}$ for each j.
- Sampling from the p-dimensional joint (posterior) distribution is often difficult.
- Computing smaller dimensional marginal (posterior) distributions is often difficult.

$$p(\theta_j) = \int p(\theta_j, \boldsymbol{\theta}_{-j}) d\boldsymbol{\theta}_{-j}$$

- Computing smaller dimensional conditional (posterior) distributions is often easy.

$$p(\theta_j \mid \boldsymbol{\theta}_{-j}) \propto p(\theta_j, \boldsymbol{\theta}_{-j}) \quad \text{with } \boldsymbol{\theta}_{-j} \text{ held constant}$$

Gibbs Sampler:

- Special type of MH sampler when the proposal is never rejected ($\alpha = 1$).
- Samples only from smaller dimensional conditional (posterior) distributions.
- **Iterative Algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  At each iteration t, for each $j = 1, \ldots, p$, sample $\theta_j^{(t)}$ from

$$p(\theta_j^{(t)} \mid \theta_1 = \theta_1^{(t)}, \cdots, \theta_{j-1} = \theta_{j-1}^{(t)}, \theta_{j+1} = \theta_{j+1}^{(t)}, \ldots, \theta_p = \theta_p^{(t-1)}).$$

Gibbs Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ and $\boldsymbol{\theta}_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)^{\mathrm{T}}$ for each j.
- Sampling from the p-dimensional joint (posterior) distribution is often difficult.
- Computing smaller dimensional marginal (posterior) distributions is often difficult.

$$p(\theta_j) = \int p(\theta_j, \boldsymbol{\theta}_{-j}) d\boldsymbol{\theta}_{-j}$$

- Computing smaller dimensional conditional (posterior) distributions is often easy.

$$p(\theta_j \mid \boldsymbol{\theta}_{-j}) \propto p(\theta_j, \boldsymbol{\theta}_{-j}) \quad \text{with } \boldsymbol{\theta}_{-j} \text{ held constant}$$

Gibbs Sampler:

- Special type of MH sampler when the proposal is never rejected ($\alpha = 1$).
- Samples only from smaller dimensional conditional (posterior) distributions.
- **Iterative Algorithm:**
  Starting with some $\boldsymbol{\theta}^{(0)}$, iteratively sample $\boldsymbol{\theta}^{(t)}$ until convergence.
  At each iteration t, for each $j = 1, \ldots, p$, sample $\theta_j^{(t)}$ from

$$p(\theta_j^{(t)} \mid \theta_1 = \theta_1^{(t)}, \cdots, \theta_{j-1} = \theta_{j-1}^{(t)}, \theta_{j+1} = \theta_{j+1}^{(t-1)}, \ldots, \theta_p = \theta_p^{(t-1)}).$$

Gibbs Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \quad \cdots$$

- $p(x, y) = \dfrac{1}{\text{Beta}(\alpha, \beta)} \dbinom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}, \quad x = 0, 1, \ldots, n, \ \ y \in (0, 1).$
    - $p(x \mid y) = \text{Bin}(n, y)$.
    - $p(y \mid x) = \text{Beta}(x + \alpha, n - x + \beta)$.
    - $n = 10, \quad \alpha = 3, \quad \beta = 3.$

Gibbs Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_{t} \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- $p(x, y) = \dfrac{1}{\text{Beta}(\alpha, \beta)} \dbinom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}, \quad x = 0, 1, \ldots, n, \;\; y \in (0, 1).$
  - $p(x \mid y) = \text{Bin}(n, y).$
  - $p(y \mid x) = \text{Beta}(x + \alpha, n - x + \beta).$
  - $n = 10, \quad \alpha = 3, \quad \beta = 3.$

Gibbs Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- $p(x, y) = \dfrac{1}{\text{Beta}(\alpha, \beta)} \dbinom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}, \quad x = 0, 1, \ldots, n, \quad y \in (0, 1).$
  - $p(x \mid y) = \text{Bin}(n, y).$
  - $p(y \mid x) = \text{Beta}(x + \alpha, n - x + \beta).$
  - $n = 10, \quad \alpha = 3, \quad \beta = 3.$

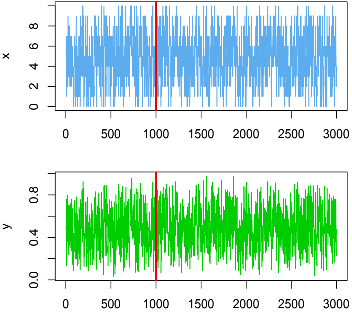
Gibbs Sampler

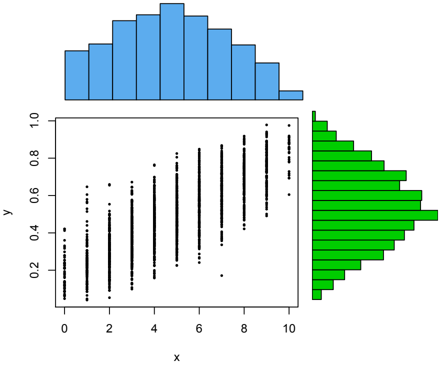
$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- $p(x, y) = \dfrac{1}{\mathrm{Beta}(\alpha, \beta)} \dbinom{n}{x} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}, \quad x = 0, 1, \ldots, n, \quad y \in (0, 1).$
  - $p(x \mid y) = \mathrm{Bin}(n, y)$.
  - $p(y \mid x) = \mathrm{Beta}(x + \alpha, n - x + \beta)$.
  - $n = 10, \quad \alpha = 3, \quad \beta = 3.$

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \to \boldsymbol{\theta}_{t-2} \to \boldsymbol{\theta}_{t-1} \to \boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1} \to \cdots$$

- $p(x, y) = \dfrac{1}{\text{Beta}(\alpha, \beta)} \dbinom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1}, \quad x = 0, 1, \ldots, n, \quad y \in (0, 1).$
  - $p(x \mid y) = \text{Bin}(n, y)$.
  - $p(y \mid x) = \text{Beta}(x + \alpha, n - x + \beta)$.
  - $n = 10, \quad \alpha = 3, \quad \beta = 3$.





Gibbs Sampler as an MH Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_{t} \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- For each $j = 1, \ldots, p$, define
$$P^G_{j,t} = p(\theta_j^\star \mid \boldsymbol{\theta}_{-j}^{(t)}, \mathbf{y}_{1:n}),$$
$$\text{where} \quad \boldsymbol{\theta}_{-j}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, \ldots, \theta_p^{(t-1)}).$$

- $P^G_{j,t} = q(\boldsymbol{\theta}_j^{(t)} \to \boldsymbol{\theta}_j^\star)$ is the proposal to move from $\boldsymbol{\theta}_j^{(t)}$ to $\boldsymbol{\theta}_j^\star$ where
$$\boldsymbol{\theta}_j^{(t)} = (\theta_1^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_j^{(t-1)}, \theta_{j+1}^{(t-1)}, \ldots, \theta_p^{(t-1)}),$$
$$\boldsymbol{\theta}_j^\star = (\theta_1^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_j^\star, \theta_{j+1}^{(t-1)}, \ldots, \theta_p^{(t-1)}).$$

- Also, $\boldsymbol{\theta}_{-j}^\star = \boldsymbol{\theta}_{-j}^{(t)}$ since other components do not change.
- Then
$$\alpha = \frac{p(\boldsymbol{\theta}_j^\star \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{P^G_{j,t}(\boldsymbol{\theta}_j^{(t)} \mid \boldsymbol{\theta}_j^\star)}{P^G_{j,t}(\boldsymbol{\theta}_j^\star \mid \boldsymbol{\theta}_j^{(t)})} = \frac{p(\boldsymbol{\theta}_j^\star \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta_j^{(t-1)} \mid \boldsymbol{\theta}_{-j}^\star, \mathbf{y}_{1:n})}{p(\theta_j^\star \mid \boldsymbol{\theta}_{-j}^{(t)}, \mathbf{y}_{1:n})}$$
$$= \frac{p(\theta_j^\star \mid \boldsymbol{\theta}_{-j}^\star, \mathbf{y}_{1:n})}{p(\theta_j^{(t-1)} \mid \boldsymbol{\theta}_{-j}^{(t)}, \mathbf{y}_{1:n})} \times \frac{p(\boldsymbol{\theta}_{-j}^\star \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_{-j}^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta_j^{(t-1)} \mid \boldsymbol{\theta}_{-j}^\star, \mathbf{y}_{1:n})}{p(\theta_j^\star \mid \boldsymbol{\theta}_{-j}^{(t)}, \mathbf{y}_{1:n})} = 1.$$

Gibbs Sampler as an MH Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- For each $j = 1, \ldots, p$, define

$$P^G_{j,t} = p(\theta^\star_j \mid \boldsymbol{\theta}^{(t)}_{-j}, \mathbf{y}_{1:n}),$$
$$\text{where} \quad \boldsymbol{\theta}^{(t)}_{-j} = (\theta^{(t)}_1, \ldots, \theta^{(t)}_{j-1}, \theta^{(t-1)}_{j+1}, \ldots, \theta^{(t-1)}_p).$$

- $P^G_{j,t} = q(\boldsymbol{\theta}^{(t)}_j \to \boldsymbol{\theta}^\star_j)$ is the proposal to move from $\boldsymbol{\theta}^{(t)}_j$ to $\boldsymbol{\theta}^\star_j$ where

$$\boldsymbol{\theta}^{(t)}_j = (\theta^{(t)}_1, \ldots, \theta^{(t)}_{j-1}, \theta^{(t-1)}_j, \theta^{(t-1)}_{j+1}, \ldots, \theta^{(t-1)}_p),$$
$$\boldsymbol{\theta}^\star_j = (\theta^{(t)}_1, \ldots, \theta^{(t)}_{j-1}, \theta^\star_j, \theta^{(t-1)}_{j+1}, \ldots, \theta^{(t-1)}_p).$$

- Also, $\boldsymbol{\theta}^\star_{-j} = \boldsymbol{\theta}^{(t)}_{-j}$ since other components do not change.
- Then

$$\alpha = \frac{p(\boldsymbol{\theta}^\star_j \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}^{(t)}_j \mid \mathbf{y}_{1:n})} \times \frac{P^G_{j,t}(\boldsymbol{\theta}^{(t)}_j \mid \boldsymbol{\theta}^\star_j)}{P^G_{j,t}(\boldsymbol{\theta}^\star_j \mid \boldsymbol{\theta}^{(t)}_j)} = \frac{p(\boldsymbol{\theta}^\star_j \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}^{(t)}_j \mid \mathbf{y}_{1:n})} \times \frac{p(\theta^{(t-1)}_j \mid \boldsymbol{\theta}^\star_{-j}, \mathbf{y}_{1:n})}{p(\theta^\star_j \mid \boldsymbol{\theta}^{(t)}_{-j}, \mathbf{y}_{1:n})}$$

$$= \frac{p(\theta^\star_j \mid \boldsymbol{\theta}^\star_{-j}, \mathbf{y}_{1:n})}{p(\theta^{(t-1)}_j \mid \boldsymbol{\theta}^{(t)}_{-j}, \mathbf{y}_{1:n})} \times \frac{p(\boldsymbol{\theta}^\star_{-j} \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}^{(t)}_{-j} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta^{(t-1)}_j \mid \boldsymbol{\theta}^\star_{-j}, \mathbf{y}_{1:n})}{p(\theta^\star_j \mid \boldsymbol{\theta}^{(t)}_{-j}, \mathbf{y}_{1:n})} = 1.$$

Gibbs Sampler as an MH Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_{t} \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- For each $j = 1, \ldots, p$, define

$$P_{j,t}^{G} = p(\theta_j^{\star} \mid \boldsymbol{\theta}_{-j}^{(t)}, \mathbf{y}_{1:n}),$$

$$\text{where} \quad \boldsymbol{\theta}_{-j}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, \ldots, \theta_p^{(t-1)}).$$

- $P_{j,t}^{G} = q(\boldsymbol{\theta}_j^{(t)} \to \boldsymbol{\theta}_j^{\star})$ is the proposal to move from $\boldsymbol{\theta}_j^{(t)}$ to $\boldsymbol{\theta}_j^{\star}$ where

$$\boldsymbol{\theta}_j^{(t)} = (\theta_1^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_j^{(t-1)}, \theta_{j+1}^{(t-1)}, \ldots, \theta_p^{(t-1)}),$$
$$\boldsymbol{\theta}_j^{\star} = (\theta_1^{(t)}, \ldots, \theta_{j-1}^{(t)}, \theta_j^{\star}, \theta_{j+1}^{(t-1)}, \ldots, \theta_p^{(t-1)}).$$

- Also, $\boldsymbol{\theta}_{-j}^{\star} = \boldsymbol{\theta}_{-j}^{(t)}$ since other components do not change.

- Then

$$\alpha = \frac{p(\boldsymbol{\theta}_j^{\star} \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{P_{j,t}^{G}(\boldsymbol{\theta}_j^{(t)} \mid \boldsymbol{\theta}_j^{\star})}{P_{j,t}^{G}(\boldsymbol{\theta}_j^{\star} \mid \boldsymbol{\theta}_j^{(t)})} = \frac{p(\boldsymbol{\theta}_j^{\star} \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta_j^{(t-1)} \mid \boldsymbol{\theta}_{-j}^{\star}, \mathbf{y}_{1:n})}{p(\theta_j^{\star} \mid \boldsymbol{\theta}_{-j}^{(t)}, \mathbf{y}_{1:n})}$$

$$= \frac{p(\theta_j^{\star} \mid \boldsymbol{\theta}_{-j}^{\star}, \mathbf{y}_{1:n})}{p(\theta_j^{(t-1)} \mid \boldsymbol{\theta}_{-j}^{(t)}, \mathbf{y}_{1:n})} \times \frac{p(\boldsymbol{\theta}_{-j}^{\star} \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_{-j}^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta_j^{(t-1)} \mid \boldsymbol{\theta}_{-j}^{\star}, \mathbf{y}_{1:n})}{p(\theta_j^{\star} \mid \boldsymbol{\theta}_{-j}^{(t)}, \mathbf{y}_{1:n})} = 1.$$

Gibbs Sampler as an MH Sampler

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

- For each $j = 1, \ldots, p$, define

$$P^G_{j,t} = p(\theta^\star_j \mid \boldsymbol{\theta}^{(t)}_{-j}, \mathbf{y}_{1:n}),$$

$$\text{where} \quad \boldsymbol{\theta}^{(t)}_{-j} = (\theta^{(t)}_1, \ldots, \theta^{(t)}_{j-1}, \theta^{(t-1)}_{j+1}, \ldots, \theta^{(t-1)}_p).$$

- $P^G_{j,t} = q(\boldsymbol{\theta}^{(t)}_j \to \boldsymbol{\theta}^\star_j)$ is the proposal to move from $\boldsymbol{\theta}^{(t)}_j$ to $\boldsymbol{\theta}^\star_j$ where

$$\boldsymbol{\theta}^{(t)}_j = (\theta^{(t)}_1, \ldots, \theta^{(t)}_{j-1}, \theta^{(t-1)}_j, \theta^{(t-1)}_{j+1}, \ldots, \theta^{(t-1)}_p),$$
$$\boldsymbol{\theta}^\star_j = (\theta^{(t)}_1, \ldots, \theta^{(t)}_{j-1}, \theta^\star_j, \theta^{(t-1)}_{j+1}, \ldots, \theta^{(t-1)}_p).$$

- Also, $\boldsymbol{\theta}^\star_{-j} = \boldsymbol{\theta}^{(t)}_{-j}$ since other components do not change.
- Then

$$\alpha = \frac{p(\boldsymbol{\theta}^\star_j \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}^{(t)}_j \mid \mathbf{y}_{1:n})} \times \frac{P^G_{j,t}(\boldsymbol{\theta}^{(t)}_j \mid \boldsymbol{\theta}^\star_j)}{P^G_{j,t}(\boldsymbol{\theta}^\star_j \mid \boldsymbol{\theta}^{(t)}_j)} = \frac{p(\boldsymbol{\theta}^\star_j \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}^{(t)}_j \mid \mathbf{y}_{1:n})} \times \frac{p(\theta^{(t-1)}_j \mid \boldsymbol{\theta}^\star_{-j}, \mathbf{y}_{1:n})}{p(\theta^\star_j \mid \boldsymbol{\theta}^{(t)}_{-j}, \mathbf{y}_{1:n})}$$

$$= \frac{p(\theta^\star_j \mid \boldsymbol{\theta}^\star_{-j}, \mathbf{y}_{1:n})}{p(\theta^{(t-1)}_j \mid \boldsymbol{\theta}^{(t)}_{-j}, \mathbf{y}_{1:n})} \times \frac{p(\boldsymbol{\theta}^\star_{-j} \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}^{(t)}_{-j} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta^{(t-1)}_j \mid \boldsymbol{\theta}^\star_{-j}, \mathbf{y}_{1:n})}{p(\theta^\star_j \mid \boldsymbol{\theta}^{(t)}_{-j}, \mathbf{y}_{1:n})} = 1.$$

Normal Model Under Conjugate Prior (From Module V)

$$y_1, \ldots, y_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2) \quad \text{with} \quad \mu, \sigma^2 \text{ both unknown}$$

- **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\overline{y}-\mu)^2\right\}\right]$
- **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}\left(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2\right)$

$$p(\mu, \sigma^2) = p(\sigma^2)p(\mu \mid \sigma^2) = \text{Inv-Ga}\left(\sigma^2 \mid \nu_0/2, \nu_0\sigma_0^2/2\right) \cdot \text{Normal}(\mu \mid \mu_0, \sigma^2/\kappa_0)$$

- **Normal-Inverse-Gamma Posterior:**

$$p(\mu, \sigma^2 \mid \mathbf{y}_{1:n})$$

$$\propto \left(\sigma^2\right)^{-\left\{\frac{(\nu_0+n)}{2}+1+\frac{1}{2}\right\}} \exp\left[-\frac{1}{2\sigma^2}\left\{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{(n+\kappa_0)}(\overline{y}-\mu_0)^2 + (\kappa_0+n)(\mu-\mu_n)^2\right\}\right]$$

$$\equiv \text{NIG}\left(\mu_n, \sigma_n^2/\kappa_n, \nu_n, \sigma_n^2\right), \quad \nu_n = (\nu_0+n), \quad \kappa_n = (\kappa_0+n), \quad \mu_n = (\kappa_0\mu_0 + n\overline{y})/(\kappa_0+n),$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left\{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{(n+\kappa_0)}(\overline{y}-\mu_0)^2\right\}$$

- **Gibbs: Iteratively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^{M}$ from the Conditional Posteriors:**
  - $p(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \text{Normal}(\mu_n, \sigma^2/\kappa_n)$
  - $p(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \text{Inv-Ga}[(\nu_n+1)/2, \{\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\}/2]$
- **Collapsed: Collectively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^{M}$ from the Marginal Posteriors:**
  - $p(\mu \mid \mathbf{y}_{1:n}) = t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$
  - $p(\sigma^2 \mid \mathbf{y}_{1:n}) = \text{Inv-Ga}(\nu_n/2, \nu_n\sigma_n^2/2)$

Normal Model Under Conjugate Prior (From Module V)

$$y_1, \ldots, y_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2) \quad \text{with} \quad \mu, \sigma^2 \text{ both unknown}$$

- **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\overline{y}-\mu)^2\right\}\right]$
- **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}\left(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2\right)$

$$p(\mu, \sigma^2) = p(\sigma^2)p(\mu \mid \sigma^2) = \text{Inv-Ga}\left(\sigma^2 \mid \nu_0/2, \nu_0\sigma_0^2/2\right) \cdot \text{Normal}(\mu \mid \mu_0, \sigma^2/\kappa_0)$$

- **Normal-Inverse-Gamma Posterior:**

$$p(\mu, \sigma^2 \mid \mathbf{y}_{1:n})$$

$$\propto \left(\sigma^2\right)^{-\left\{\frac{(\nu_0+n)}{2}+1+\frac{1}{2}\right\}} \exp\left[-\frac{1}{2\sigma^2}\left\{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{(n+\kappa_0)}(\overline{y}-\mu_0)^2 + (\kappa_0+n)(\mu-\mu_n)^2\right\}\right]$$

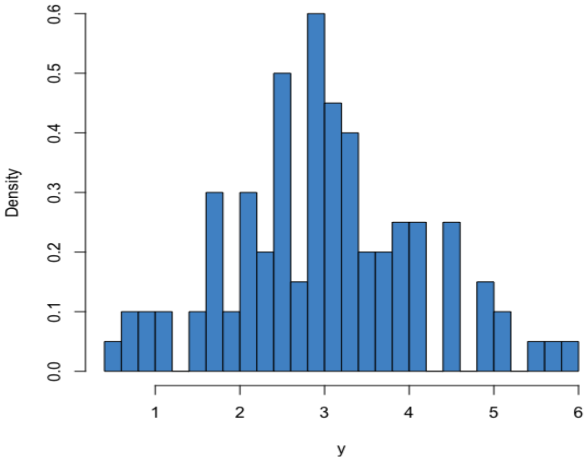
$$\equiv \text{NIG}\left(\mu_n, \sigma_n^2/\kappa_n, \nu_n, \sigma_n^2\right), \quad \nu_n = (\nu_0+n), \quad \kappa_n = (\kappa_0+n), \quad \mu_n = (\kappa_0\mu_0 + n\overline{y})/(\kappa_0+n),$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left\{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{(n+\kappa_0)}(\overline{y}-\mu_0)^2\right\}$$

- **Gibbs: Iteratively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^M$ from the Conditional Posteriors:**
  - $p(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \text{Normal}(\mu_n, \sigma^2/\kappa_n)$
  - $p(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \text{Inv-Ga}[(\nu_n+1)/2, \{\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\}/2]$
- **Collapsed: Collectively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^M$ from the Marginal Posteriors:**
  - $p(\mu \mid \mathbf{y}_{1:n}) = t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$
  - $p(\sigma^2 \mid \mathbf{y}_{1:n}) = \text{Inv-Ga}(\nu_n/2, \nu_n\sigma_n^2/2)$

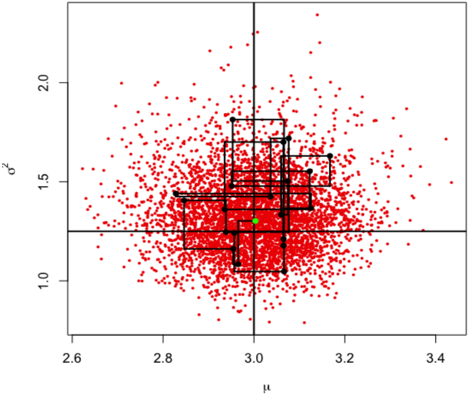
Gibbs Sampler for Normal Model Under NIG Prior - Data

$$y_1, \ldots, y_{100} \overset{iid}{\sim} \text{Normal}(3, \sqrt{1.25}^2) \equiv \text{Normal}(3, 1.25)$$
$$(\mu, \sigma^2) \sim \text{NIG}\left(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2\right).$$



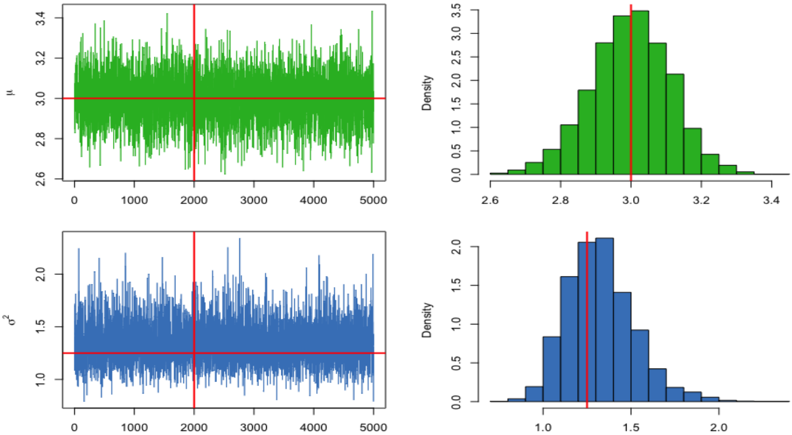
Gibbs Sampler for Normal Model Under NIG Prior - Sample Path

$$y_1, \ldots, y_{100} \overset{iid}{\sim} \text{Normal}(3, \sqrt{1.25}^2) \equiv \text{Normal}(3, 1.25)$$
$$(\mu, \sigma^2) \sim \text{NIG}\left(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2\right).$$



Gibbs Sampler for Normal Model Under NIG Prior - Trace Plots

$$y_1, \ldots, y_{100} \overset{iid}{\sim} \text{Normal}(3, \sqrt{1.25}^2) \equiv \text{Normal}(3, 1.25)$$
$$(\mu, \sigma^2) \sim \text{NIG}\left(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2\right).$$



Gibbs Sampler - Rao-Blackwellization

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_{t} \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

$$\mathbb{E}_x(x) = \mathbb{E}_y(\mathbb{E}_{x|y}(x \mid y)), \qquad \text{var}_x(x) = \text{var}_y(\mathbb{E}_{x|y}(x \mid y)) + \mathbb{E}_y(\text{var}_{x|y}(x \mid y))$$

Moments:

- Moments of x can be estimated straightforwardly using Gibbs samples $x^{(t)}$ as

$$\mathbb{E}(x) \mathrel{\widehat{=}} \frac{1}{T}\sum_{t=1}^{T} x^{(t)}.$$

- An alternative estimate constructed using Gibbs samples $y^{(t)}$ is

$$\mathbb{E}\mathbb{E}(x \mid y) = \mathbb{E}\{g(y)\} \mathrel{\widehat{=}} \frac{1}{T}\sum_{t=1}^{T} g\left(y^{(t)}\right).$$

- Marginal of x can be estimated straightforwardly using Gibbs samples $x^{(t)}$.

- An alternative estimate that often better estimates the uses Gibbs samples $y^{(t)}$ as

$$p(x) = \int p(x \mid y)p(y)dy = \mathbb{E}_{y\sim p(y)} p(x \mid y) \mathrel{\widehat{=}} \frac{1}{T}\sum_{t=1}^{T} p(x \mid y^{(t)}).$$

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_{t} \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

$$\mathbb{E}_x(x) = \mathbb{E}_y(\mathbb{E}_{x|y}(x \mid y)), \qquad \mathrm{var}_x(x) = \mathrm{var}_y(\mathbb{E}_{x|y}(x \mid y)) + \mathbb{E}_y(\mathrm{var}_{x|y}(x \mid y))$$

Moments:

- Moments of x can be estimated straightforwardly using Gibbs samples $x^{(t)}$ as

$$\mathbb{E}(x) \mathrel{\widehat{=}} \frac{1}{T}\sum_{t=1}^{T} x^{(t)}.$$

- An alternative estimate constructed using Gibbs samples $y^{(t)}$ is

$$\mathbb{E}\mathbb{E}(x \mid y) = \mathbb{E}\{g(y)\} \mathrel{\widehat{=}} \frac{1}{T}\sum_{t=1}^{T} g\left(y^{(t)}\right).$$

- Marginal of x can be estimated straightforwardly using Gibbs samples $x^{(t)}$.

- An alternative estimate that often better estimates the uses Gibbs samples $y^{(t)}$ as

$$p(x) = \int p(x \mid y)p(y)dy = \mathbb{E}_{y\sim p(y)}p(x \mid y) \mathrel{\widehat{=}} \frac{1}{T}\sum_{t=1}^{T} p(x \mid y^{(t)}).$$

Gibbs Sampler - Rao-Blackwellization

$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t)$

$$\cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_{t} \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

$$\mathbb{E}_x(x) = \mathbb{E}_y(\mathbb{E}_{x|y}(x \mid y)), \qquad \mathrm{var}_x(x) = \mathrm{var}_y(\mathbb{E}_{x|y}(x \mid y)) + \mathbb{E}_y(\mathrm{var}_{x|y}(x \mid y))$$

Moments:

- Moments of x can be estimated straightforwardly using Gibbs samples $x^{(t)}$ as

$$\mathbb{E}(x) \mathrel{\widehat{=}} \frac{1}{T}\sum_{t=1}^{T} x^{(t)}.$$

- An alternative estimate constructed using Gibbs samples $y^{(t)}$ is

$$\mathbb{E}\mathbb{E}(x \mid y) = \mathbb{E}\{g(y)\} \mathrel{\widehat{=}} \frac{1}{T}\sum_{t=1}^{T} g\left(y^{(t)}\right).$$

Marginal Distributions:

- Marginal of x can be estimated straightforwardly using Gibbs samples $x^{(t)}$.

- An alternative estimate that often better estimates the uses Gibbs samples $y^{(t)}$ as

$$p(x) = \int p(x \mid y)p(y)dy = \mathbb{E}_{y\sim p(y)}p(x \mid y) \mathrel{\widehat{=}} \frac{1}{T}\sum_{t=1}^{T} p(x \mid y^{(t)}).$$

Gibbs Sampler - Rao-Blackwellization

$$p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_t) = p(\boldsymbol{\theta}_{t+1} \mid \boldsymbol{\theta}_t) \qquad \cdots \longrightarrow \boldsymbol{\theta}_{t-2} \longrightarrow \boldsymbol{\theta}_{t-1} \longrightarrow \boldsymbol{\theta}_t \longrightarrow \boldsymbol{\theta}_{t+1} \longrightarrow \cdots$$

$$\mathbb{E}_x(x) = \mathbb{E}_y(\mathbb{E}_{x|y}(x \mid y)), \qquad \mathrm{var}_x(x) = \mathrm{var}_y(\mathbb{E}_{x|y}(x \mid y)) + \mathbb{E}_y(\mathrm{var}_{x|y}(x \mid y))$$

Moments:

- Moments of x can be estimated straightforwardly using Gibbs samples $x^{(t)}$ as

$$\mathbb{E}(x) \widehat{=} \frac{1}{T} \sum_{t=1}^{T} x^{(t)}.$$

- An alternative estimate constructed using Gibbs samples $y^{(t)}$ is

$$\mathbb{E}\mathbb{E}(x \mid y) = \mathbb{E}\{g(y)\} \widehat{=} \frac{1}{T} \sum_{t=1}^{T} g\left(y^{(t)}\right).$$

Marginal Distributions:

- Marginal of x can be estimated straightforwardly using Gibbs samples $x^{(t)}$.

- An alternative estimate that often better estimates the uses Gibbs samples $y^{(t)}$ as

$$p(x) = \int p(x \mid y)p(y)dy = \mathbb{E}_{y \sim p(y)} p(x \mid y) \widehat{=} \frac{1}{T} \sum_{t=1}^{T} p(x \mid y^{(t)}).$$

Normal Model Under NIG Prior - Rao-Blackwellization

$$y_1, \ldots, y_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2) \quad \text{with} \quad \mu, \sigma^2 \text{ both unknown}$$

- **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\overline{y}-\mu)^2\right\}\right]$
- **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}\left(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2\right)$

$$p(\mu, \sigma^2) = p(\sigma^2)p(\mu \mid \sigma^2) = \text{Inv-Ga}\left(\sigma^2 \mid \nu_0/2, \nu_0\sigma_0^2/2\right) \cdot \text{Normal}(\mu \mid \mu_0, \sigma^2/\kappa_0)$$

- **Normal-Inverse-Gamma Posterior:**

$$p(\mu, \sigma^2 \mid \mathbf{y}_{1:n}) = \text{NIG}\left(\mu_n, \sigma_n^2/\kappa_n, \nu_n, \sigma_n^2\right), \quad \nu_n = (\nu_0 + n), \quad \kappa_n = (\kappa_0 + n), \quad \mu_n = (\kappa_0\mu_0 + n\overline{y})/(\kappa_0 + n),$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left\{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{(n+\kappa_0)}(\overline{y}-\mu_0)^2\right\}$$

- **Gibbs: Iteratively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^M$ from the Conditional Posteriors:**
  - $p(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \text{Normal}(\mu_n, \sigma^2/\kappa_n)$ $\quad \Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^M \mu^{(m)}$
  - $p(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \text{Inv-Ga}[(\nu_n+1)/2, \{\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\}/2]$ $\quad \Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^M \sigma^{2(m)}$
- **Collapsed: Collectively Sample from the Marginal Posteriors:**
  - $p(\mu \mid \mathbf{y}_{1:n}) = t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$ $\quad \Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^M \mu^{(m)}$
  - $p(\sigma^2 \mid \mathbf{y}_{1:n}) = \text{Inv-Ga}(\nu_n/2, \nu_n\sigma_n^2/2)$ $\quad \Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^M \sigma^{2(m)}$
- **Rao-Blackwellized: Iteratively Sample from the Conditional Posteriors:**
  - $\mathbb{E}(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \mu_{n|\sigma^2} = \mu_n = (\kappa_0\mu_0 + n\overline{y})/(\kappa_0 + n)$
  - $\mathbb{E}(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \sigma_{n|\mu}^2 = \{\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\}/(\nu_n - 1)$

Normal Model Under NIG Prior - Rao-Blackwellization

$$y_1, \ldots, y_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2) \quad \text{with} \quad \mu, \sigma^2 \text{ both unknown}$$

- **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\overline{y}-\mu)^2\right\}\right]$
- **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}\left(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2\right)$

$$p(\mu, \sigma^2) = p(\sigma^2)p(\mu \mid \sigma^2) = \text{Inv-Ga}\left(\sigma^2 \mid \nu_0/2, \nu_0\sigma_0^2/2\right) \cdot \text{Normal}(\mu \mid \mu_0, \sigma^2/\kappa_0)$$

- **Normal-Inverse-Gamma Posterior:**

$$p(\mu, \sigma^2 \mid \mathbf{y}_{1:n}) = \text{NIG}\left(\mu_n, \sigma_n^2/\kappa_n, \nu_n, \sigma_n^2\right), \quad \nu_n = (\nu_0 + n), \quad \kappa_n = (\kappa_0 + n), \quad \mu_n = (\kappa_0\mu_0 + n\overline{y})/(\kappa_0 + n),$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left\{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{(n+\kappa_0)}(\overline{y}-\mu_0)^2\right\}$$

- **Gibbs: Iteratively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^{M}$ from the Conditional Posteriors:**
  - $p(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \text{Normal}(\mu_n, \sigma^2/\kappa_n)$ $\Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^{M}\mu^{(m)}$
  - $p(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \text{Inv-Ga}[(\nu_n+1)/2, \{\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\}/2]$ $\Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^{M}\sigma^{2(m)}$
- **Collapsed: Collectively Sample from the Marginal Posteriors:**
  - $p(\mu \mid \mathbf{y}_{1:n}) = t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$ $\Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^{M}\mu^{(m)}$
  - $p(\sigma^2 \mid \mathbf{y}_{1:n}) = \text{Inv-Ga}(\nu_n/2, \nu_n\sigma_n^2/2)$ $\Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^{M}\sigma^{2(m)}$
- **Rao-Blackwellized: Iteratively Sample from the Conditional Posteriors:**
  - $\mathbb{E}(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \mu_{n|\sigma^2} = \mu_n = (\kappa_0\mu_0 + n\overline{y})/(\kappa_0 + n)$
  - $\mathbb{E}(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \sigma^2_{n|\mu} = \{\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\}/(\nu_n - 1)$

Normal Model Under NIG Prior - Rao-Blackwellization

$$y_1, \ldots, y_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2) \quad \text{with} \quad \mu, \sigma^2 \text{ both unknown}$$

- **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\overline{y}-\mu)^2\right\}\right]$
- **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}\left(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2\right)$

$$p(\mu, \sigma^2) = p(\sigma^2)p(\mu \mid \sigma^2) = \text{Inv-Ga}\left(\sigma^2 \mid \nu_0/2, \nu_0\sigma_0^2/2\right) \cdot \text{Normal}(\mu \mid \mu_0, \sigma^2/\kappa_0)$$

- **Normal-Inverse-Gamma Posterior:**

$$p(\mu, \sigma^2 \mid \mathbf{y}_{1:n}) = \text{NIG}\left(\mu_n, \sigma_n^2/\kappa_n, \nu_n, \sigma_n^2\right), \quad \nu_n = (\nu_0 + n), \quad \kappa_n = (\kappa_0 + n), \quad \mu_n = (\kappa_0\mu_0 + n\overline{y})/(\kappa_0 + n),$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left\{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{(n+\kappa_0)}(\overline{y}-\mu_0)^2\right\}$$

- **Gibbs: Iteratively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^{M}$ from the Conditional Posteriors:**
  - $p(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \text{Normal}(\mu_n, \sigma^2/\kappa_n)$ $\quad\Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^{M}\mu^{(m)}$
  - $p(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \text{Inv-Ga}[(\nu_n+1)/2, \{\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\}/2]$ $\quad\Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^{M}\sigma^{2(m)}$
- **Collapsed: Collectively Sample from the Marginal Posteriors:**
  - $p(\mu \mid \mathbf{y}_{1:n}) = t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$ $\quad\Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^{M}\mu^{(m)}$
  - $p(\sigma^2 \mid \mathbf{y}_{1:n}) = \text{Inv-Ga}(\nu_n/2, \nu_n\sigma_n^2/2)$ $\quad\Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^{M}\sigma^{2(m)}$
- **Rao-Blackwellized: Iteratively Sample from the Conditional Posteriors:**
  - $\mathbb{E}(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \mu_{n\mid\sigma^2} = \mu_n = (\kappa_0\mu_0 + n\overline{y})/(\kappa_0 + n)$ $\quad\Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \widehat{=} \mu_n$
  - $\mathbb{E}(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \sigma_{n\mid\mu}^2 = \{\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\}/(\nu_n - 1)$ $\quad\Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^{M}\sigma_{n\mid\mu}^{2(m)}$

Normal Model Under NIG Prior - Rao-Blackwellization

$$y_1, \ldots, y_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2) \quad \text{with} \quad \mu, \sigma^2 \text{ both unknown}$$

- **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left\{(n-1)s^2 + n(\overline{y}-\mu)^2\right\}\right]$
- **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}\left(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2\right)$

$$p(\mu, \sigma^2) = p(\sigma^2)p(\mu \mid \sigma^2) = \text{Inv-Ga}\left(\sigma^2 \mid \nu_0/2, \nu_0\sigma_0^2/2\right) \cdot \text{Normal}(\mu \mid \mu_0, \sigma^2/\kappa_0)$$

- **Normal-Inverse-Gamma Posterior:**

$$p(\mu, \sigma^2 \mid \mathbf{y}_{1:n}) = \text{NIG}\left(\mu_n, \sigma_n^2/\kappa_n, \nu_n, \sigma_n^2\right), \quad \nu_n = (\nu_0 + n), \quad \kappa_n = (\kappa_0 + n), \quad \mu_n = (\kappa_0\mu_0 + n\overline{y})/(\kappa_0 + n),$$

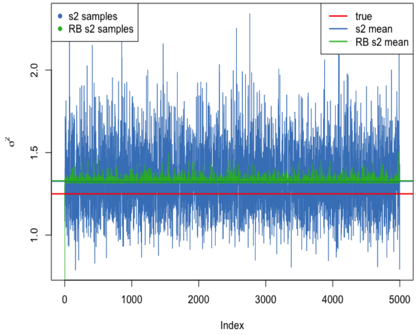
$$\sigma_n^2 = \frac{1}{\nu_n}\left\{\nu_0\sigma_0^2 + (n-1)s^2 + \frac{n\kappa_0}{(n+\kappa_0)}(\overline{y}-\mu_0)^2\right\}$$

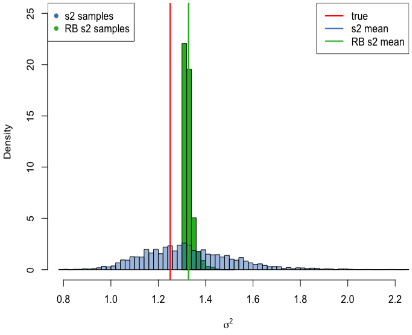
- **Gibbs: Iteratively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^M$ from the Conditional Posteriors:**
  - $p(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \text{Normal}(\mu_n, \sigma^2/\kappa_n)$ $\quad\Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^M \mu^{(m)}$
  - $p(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \text{Inv-Ga}[(\nu_n + 1)/2, \{\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\}/2]$ $\quad\Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^M \sigma^{2(m)}$
- **Collapsed: Collectively Sample from the Marginal Posteriors:**
  - $p(\mu \mid \mathbf{y}_{1:n}) = t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$ $\quad\Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^M \mu^{(m)}$
  - $p(\sigma^2 \mid \mathbf{y}_{1:n}) = \text{Inv-Ga}(\nu_n/2, \nu_n\sigma_n^2/2)$ $\quad\Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \widehat{=} \frac{1}{M}\sum_{m=1}^M \sigma^{2(m)}$
- **Rao-Blackwellized: Iteratively Sample from the Conditional Posteriors:**
  - $\mathbb{E}(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \mu_{n|\sigma^2} = \mu_n = (\kappa_0\mu_0 + n\overline{y})/(\kappa_0 + n)$ $\rightarrow$ **Does not depend on the samples at all!**
  - $\mathbb{E}(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \sigma_{n|\mu}^2 = \{\kappa_n(\mu-\mu_n)^2 + \nu_n\sigma_n^2\}/(\nu_n - 1)$ $\rightarrow$ **Depends only on the samples of μ!**

Gibbs Sampler for Normal Model Under NIG Prior - Rao-Blackwellization

$$y_1, \ldots, y_{100} \overset{iid}{\sim} \text{Normal}(3, \sqrt{1.25}^2) \equiv \text{Normal}(3, 1.25)$$
$$(\mu, \sigma^2) \sim \text{NIG}\left(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2\right).$$

$$\widehat{\sigma}^2_{post-mean} = \frac{1}{M} \sum_{m=1}^{M} \sigma^{2(m)}$$
$$\widehat{\sigma}^2_{RB-post-mean} = \frac{1}{M} \sum_{m=1}^{M} \frac{\kappa_n(\mu^{(m)} - \mu_n)^2 + \nu_n \sigma_n^2}{\nu_n - 1}$$





Module VII Summary

- Monte Carlo integration
  - Approximates integrals writing it as an expectation then sampling from the corresponding density
  - Importance sampling
    - Allows Monte Carlo integration when sampling from the target density is difficult
    - Uses weighted samples from a reference density that is easy to sample
    - Applicable even when the target density is known only up to a normalizing constant
- Sampling from a target density when direct sampling is difficult
  - Importance resampling
  - Markov chain Monte Carlo

Module VII Summary

- Monte Carlo integration
  - Approximates integrals writing it as an expectation then sampling from the corresponding density
  - Importance sampling
    - Allows Monte Carlo integration when sampling from the target density is difficult
    - Uses weighted samples from a reference density that is easy to sample
    - Applicable even when the target density is known only up to a normalizing constant
- Sampling from a target density when direct sampling is difficult
  - Importance resampling
  - Markov chain Monte Carlo

- Monte Carlo integration
  - Approximates integrals writing it as an expectation then sampling from the corresponding density
  - Importance sampling
    - Allows Monte Carlo integration when sampling from the target density is difficult
    - Uses weighted samples from a reference density that is easy to sample
    - Applicable even when the target density is known only up to a normalizing constant
- Sampling from a target density when direct sampling is difficult
  - Importance resampling
  - Markov chain Monte Carlo

Module VII Summary

- Monte Carlo integration
  - Approximates integrals writing it as an expectation then sampling from the corresponding density
  - Importance sampling
    - Allows Monte Carlo integration when sampling from the target density is difficult
    - Uses weighted samples from a reference density that is easy to sample
    - Applicable even when the target density is known only up to a normalizing constant
- Sampling from a target density when direct sampling is difficult
  - Importance resampling
  - Markov chain Monte Carlo

Module VII Summary

- Monte Carlo integration
  - Approximates integrals writing it as an expectation then sampling from the corresponding density
  - Importance sampling
    - Allows Monte Carlo integration when sampling from the target density is difficult
    - Uses weighted samples from a reference density that is easy to sample
    - Applicable even when the target density is known only up to a normalizing constant
- Sampling from a target density when direct sampling is difficult
  - Importance resampling
  - Markov chain Monte Carlo

Module VII Summary

- Monte Carlo integration
  - Approximates integrals writing it as an expectation then sampling from the corresponding density
  - Importance sampling
    - Allows Monte Carlo integration when sampling from the target density is difficult
    - Uses weighted samples from a reference density that is easy to sample
    - Applicable even when the target density is known only up to a normalizing constant
- Sampling from a target density when direct sampling is difficult
  - Importance sampling
  - Markov chain Monte Carlo

Module VII Summary

- Monte Carlo integration
  - Approximates integrals writing it as an expectation then sampling from the corresponding density
  - Importance sampling
    - Allows Monte Carlo integration when sampling from the target density is difficult
    - Uses weighted samples from a reference density that is easy to sample
    - Applicable even when the target density is known only up to a normalizing constant
- Sampling from a target density when direct sampling is difficult
  - Importance resampling
  - Markov chain Monte Carlo

Module VII Summary

- Monte Carlo integration
  - Approximates integrals writing it as an expectation then sampling from the corresponding density
  - Importance sampling
    - Allows Monte Carlo integration when sampling from the target density is difficult
    - Uses weighted samples from a reference density that is easy to sample
    - Applicable even when the target density is known only up to a normalizing constant
- Sampling from a target density when direct sampling is difficult
  - Importance resampling
    - Resamples draws from a reference density with importance sampling weights
  - Markov chain Monte Carlo
    - Samples from a stationary Markov chain with the target density as the stationary distribution

Module VII Summary

- Monte Carlo integration
  - Approximates integrals writing it as an expectation then sampling from the corresponding density
  - Importance sampling
    - Allows Monte Carlo integration when sampling from the target density is difficult
    - Uses weighted samples from a reference density that is easy to sample
    - Applicable even when the target density is known only up to a normalizing constant
- Sampling from a target density when direct sampling is difficult
  - Importance resampling
    - Resamples draws from a reference density with importance sampling weights
  - Markov chain Monte Carlo

Module VII Summary

- Monte Carlo integration
  - Approximates integrals writing it as an expectation then sampling from the corresponding density
  - Importance sampling
    - Allows Monte Carlo integration when sampling from the target density is difficult
    - Uses weighted samples from a reference density that is easy to sample
    - Applicable even when the target density is known only up to a normalizing constant
- Sampling from a target density when direct sampling is difficult
  - Importance resampling
    - Resamples draws from a reference density with importance sampling weights
  - Markov chain Monte Carlo
    - Samples from a stationary Markov chain with the target density as the stationary distribution

Module VII Summary

- Monte Carlo integration
  - Approximates integrals writing it as an expectation then sampling from the corresponding density
  - Importance sampling
    - Allows Monte Carlo integration when sampling from the target density is difficult
    - Uses weighted samples from a reference density that is easy to sample
    - Applicable even when the target density is known only up to a normalizing constant
- Sampling from a target density when direct sampling is difficult
  - Importance resampling
    - Resamples draws from a reference density with importance sampling weights
  - Markov chain Monte Carlo
    - Samples from a stationary Markov chain with the target density as the stationary distribution

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
  - Metropolis-Hastings sampler
  - Gibbs sampler
  - Convergence diagnostics

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
  - Gibbs sampler
  - Convergence diagnostics

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
  - Gibbs sampler
  - Convergence diagnostics

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
  - Gibbs sampler
  - Convergence diagnostics

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
  - Convergence diagnostics

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
  - Convergence diagnostics

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning

Module VII Summary Continued

- Markov chains
  - The two components that define an MC are
    - Initial distribution $p(z_1)$
    - Transition distributions $p(z_t \mid z_{t-1})$
  - Important properties
    - Irreducibility - all states communicate with each other
    - Aperiodicity - no cycles of fixed lengths between states
    - Stationarity - probability of visiting a state is independent of the initial state
    - Reversibility - probabilities of moves forward and backward between any two states are equal
  - Metropolis-Hastings sampler
    - Used to sample from a target distribution that is difficult to directly sample from
    - Constructs a Markov chain with the target distribution as the stationary distribution
    - At each iteration, accepts or rejects proposed values with certain probabilities
    - Performance depends heavily on the choice of proposal distributions
    - Not suitable for sampling from complex, multi-dimensional distributions
  - Gibbs sampler
    - A special type of MH algorithm that always accepts
    - Suitable for sampling from complex, multi-dimensional distributions
    - Samples iteratively from conditional distributions of each variable given all others
  - Convergence diagnostics
    - Trace plots and autocorrelation plots
    - Burn-in and thinning