

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

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y}) d\mathbf{y}$, $\mathcal{Y} \subseteq \mathbb{R}^d$.

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

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

- Monte Carlo standard error

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

- 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

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

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y}) d\mathbf{y}$, $\mathcal{Y} \subseteq \mathbb{R}^d$.
- 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 $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $p(\mathbf{y})$. Then

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

- Monte Carlo standard error

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

- 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

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

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y}) d\mathbf{y}$, $\mathcal{Y} \subseteq \mathbb{R}^d$.
- 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 $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $p(\mathbf{y})$. Then

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

- Monte Carlo standard error

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

- 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

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

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y}) d\mathbf{y}$, $\mathcal{Y} \subseteq \mathbb{R}^d$.
- 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 $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $p(\mathbf{y})$. Then

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

- Monte Carlo standard error

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

- 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

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

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y}) d\mathbf{y}$, $\mathcal{Y} \subseteq \mathbb{R}^d$.
- 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 $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $p(\mathbf{y})$. Then

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

- Monte Carlo standard error

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

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

- 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

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

- We want to compute the integral $\int_{\mathcal{Y}} h(\mathbf{y}) d\mathbf{y}$, $\mathcal{Y} \subseteq \mathbb{R}^d$.
- 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 $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $p(\mathbf{y})$. Then

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

- Monte Carlo standard error

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

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

- 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

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

- 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 $\mathbf{I} = \int_{\mathbf{y}} g(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} = \int_{\mathbf{y}} g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} q(\mathbf{y}) d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\}$.
- Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

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

- An alternative formulation is

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

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

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

- $\hat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\hat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \rightarrow \infty$.
- $\hat{\mathbf{I}}_2$ 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 $\mathbf{I} = \int_{\mathbf{y}} g(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} = \int_{\mathbf{y}} g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} q(\mathbf{y}) d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\}$.

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

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

- An alternative formulation is

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

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

- Monte Carlo standard error

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

- $\hat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\hat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \rightarrow \infty$.

- $\hat{\mathbf{I}}_2$ 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 $\mathbf{I} = \int_{\mathbf{y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathbf{y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\}$.
- Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

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

- An alternative formulation is

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

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

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

- $\hat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\hat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \rightarrow \infty$.
- $\hat{\mathbf{I}}_2$ 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 $\mathbf{I} = \int_{\mathbf{y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathbf{y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\}$.
- Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

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

- An alternative formulation is

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

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

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

- $\hat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\hat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \rightarrow \infty$.
- $\hat{\mathbf{I}}_2$ 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 $\mathbf{I} = \int_{\mathbf{y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathbf{y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\}$.
- Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

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

- An alternative formulation is

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

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

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

- $\hat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\hat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \rightarrow \infty$.
- $\hat{\mathbf{I}}_2$ 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 $\mathbf{I} = \int_{\mathbf{y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathbf{y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\}$.
- Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

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

- An alternative formulation is

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

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

- Monte Carlo standard error

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

- $\hat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\hat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \rightarrow \infty$.
- $\hat{\mathbf{I}}_2$ 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 $\mathbf{I} = \int_{\mathbf{y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathbf{y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\}$.
- Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

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

- An alternative formulation is

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

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

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

- $\hat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\hat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \rightarrow \infty$.
- $\hat{\mathbf{I}}_2$ 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 $\mathbf{I} = \int_{\mathbf{y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathbf{y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\}$.
- Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

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

- An alternative formulation is

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

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

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

- $\hat{\mathbf{I}}_1 \xrightarrow{P} \mathbf{I}$ and $\hat{\mathbf{I}}_2 \xrightarrow{P} \mathbf{I}$ as $n \rightarrow \infty$.

• $\hat{\mathbf{I}}_2$ 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 $\mathbf{I} = \int_{\mathbf{y}} g(\mathbf{y})p(\mathbf{y})d\mathbf{y} = \int_{\mathbf{y}} g(\mathbf{y})\frac{p(\mathbf{y})}{q(\mathbf{y})}q(\mathbf{y})d\mathbf{y} = \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y})} \left\{ g(\mathbf{y}) \frac{p(\mathbf{y})}{q(\mathbf{y})} \right\}$.
- Let $\mathbf{y}_1, \dots, \mathbf{y}_n$ be a random sample from $q(\mathbf{y})$. Then

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

- An alternative formulation is

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

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

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

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

- ▶ Likelihood: $y_1, \dots, y_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$ with $p(y_i | \theta) = \theta^{y_i} (1 - \theta)^{1-y_i}$
- ▶ Prior: $p(\theta) \propto \cos^2(4\pi\theta) = \tilde{p}(\theta)$
- ▶ Posterior: $p(\theta | \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta) \theta^s (1 - \theta)^{n-s} = \tilde{p}(\theta | \mathbf{y}_{1:n})$

- ▶ Likelihood: $y_1, \dots, y_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$ with $p(y_i | \theta) = \theta^{y_i} (1 - \theta)^{1-y_i}$
- ▶ Prior: $p(\theta) \propto \cos^2(4\pi\theta) = \tilde{p}(\theta)$
- ▶ Posterior: $p(\theta | \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta) \theta^s (1 - \theta)^{n-s} = \tilde{p}(\theta | \mathbf{y}_{1:n})$

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

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

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

Importance Sampling

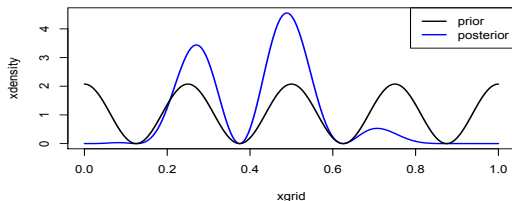
- ▶ Likelihood: $y_1, \dots, y_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$ with $p(y_i | \theta) = \theta^{y_i} (1 - \theta)^{1-y_i}$
- ▶ Prior: $p(\theta) \propto \cos^2(4\pi\theta) = \tilde{p}(\theta)$
- ▶ Posterior: $p(\theta | \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta) \theta^s (1 - \theta)^{n-s} = \tilde{p}(\theta | \mathbf{y}_{1:n})$

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

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

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

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



- 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

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

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

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

- An alternative formulation is

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

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

- 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

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

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

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

- An alternative formulation is

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

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

- 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

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

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

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

- An alternative formulation is

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

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

- 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

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

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

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

- An alternative formulation is

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

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

- 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

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

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

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

- An alternative formulation is

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

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

- 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

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

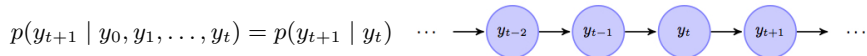
- Let z_1, \dots, z_n be a random sample from $q(z)$. Then

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

- An alternative formulation is

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

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



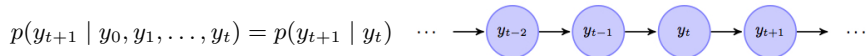
- **Initial Distribution:** probability that the chain starts with a state y :

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

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

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

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



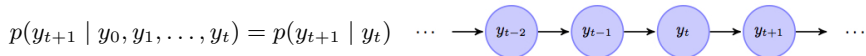
- **Initial Distribution:** probability that the chain starts with a state y :

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

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

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

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



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

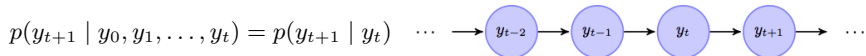
$$p(y_0 = y) = \pi_0(y).$$

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

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

- **Transition Probability Matrix:** $\mathbf{P} = \left(\left(p(x, y) \right) \right)$.

- The probability that the chain is in state y at time t : $p(y_t = y) = \pi_t(y)$.



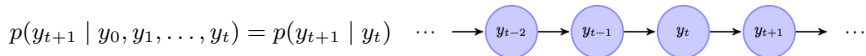
- **Initial Distribution:** probability that the chain starts with a state y :

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

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

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

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



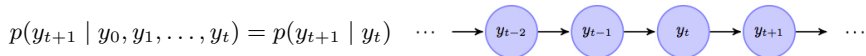
- **Initial Distribution:** probability that the chain starts with a state y :

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

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

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

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



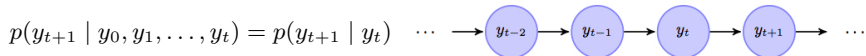
- **Initial Distribution:** probability that the chain starts with a state y :

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

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

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

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



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

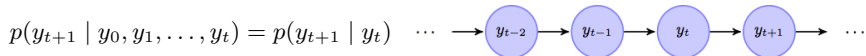
$$p(y_0 = y) = \pi_0(y).$$

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

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

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

- **Chapman-Kolmogorov Equation:** $\pi_{t+1}(y) = p(y_{t+1} = y)$
 $= \sum_{x \in \mathcal{Y}} p(y_{t+1} = y \mid y_t = x) p(y_t = x)$
 $= \sum_{x \in \mathcal{Y}} p(x, y) \pi_t(x).$



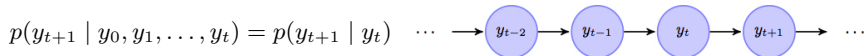
- **Initial Distribution:** probability that the chain starts with a state y :

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

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

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

- **Transition Probability Matrix:** $\mathbf{P} = \left(\left(p(x, y) \right) \right)$.
- The probability that the chain is in state y at time t : $p(y_t = y) = \pi_t(y)$.
- We have $\mathbf{P}\mathbf{1} = \mathbf{1}$, $\pi_t\mathbf{1} = 1$.
- **Chapman-Kolmogorov Equation:** $\pi_{t+1} = \pi_t\mathbf{P}$.



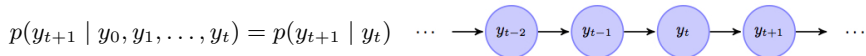
- **Initial Distribution:** probability that the chain starts with a state y :

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

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

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

- **Transition Probability Matrix:** $\mathbf{P} = \left(\left(p(x, y) \right) \right)$.
- The probability that the chain is in state y at time t : $p(y_t = y) = \pi_t(y)$.
- We have $\mathbf{P}\mathbf{1} = \mathbf{1}$, $\pi_t\mathbf{1} = 1$.
- **Chapman-Kolmogorov Equation:** $\pi_{t+1} = \pi_t\mathbf{P}$.
- We have $\pi_t = \pi_{t-1}\mathbf{P} = \pi_{t-2}\mathbf{P}^2 = \dots = \pi_0\mathbf{P}^t$.



- **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

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

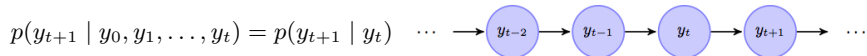
- **Aperiodicity:** A Markov chain is aperiodic if the number of steps required to move between two states is not a multiple of some integer:

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

- 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

$$\pi = \pi P.$$

- Irreducible and aperiodic Markov chains converge to stationarity.



- **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:

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

- 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

$$\pi = \pi P.$$

- Irreducible and aperiodic Markov chains converge to stationarity.

$$p(y_{t+1} \mid y_0, y_1, \dots, y_t) = p(y_{t+1} \mid y_t) \quad \cdots \rightarrow \textcircled{y_{t-2}} \rightarrow \textcircled{y_{t-1}} \rightarrow \textcircled{y_t} \rightarrow \textcircled{y_{t+1}} \rightarrow \cdots$$

- **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

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

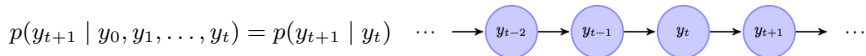
- **Aperiodicity:** A Markov chain is aperiodic if the number of steps required to move between two states is not a multiple of some integer:

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

- 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

$$\pi = \pi P.$$

- Irreducible and aperiodic Markov chains converge to stationarity.



- **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:

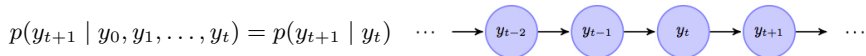
$$\text{GCD}\{n : p^n(x, y) > 0 \quad \forall x, y \in \mathcal{Y}\} = 1.$$

- 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

$$\pi = \pi P.$$

- Irreducible and aperiodic Markov chains converge to stationarity.



- **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

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

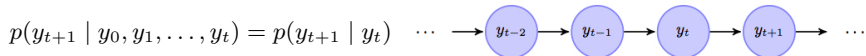
- **Aperiodicity:** A Markov chain is aperiodic if the number of steps required to move between two states is not a multiple of some integer:

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

- 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

$$\pi = \pi P.$$

- Irreducible and aperiodic Markov chains converge to stationarity.



- **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:

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

- 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.

$$p(y_{t+1} \mid y_0, y_1, \dots, y_t) = p(y_{t+1} \mid y_t) \quad \cdots \rightarrow y_{t-2} \rightarrow y_{t-1} \rightarrow y_t \rightarrow y_{t+1} \rightarrow \cdots$$

► $\mathbf{P} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0.4 & 0.4 & 0.2 \\ 0 & 0 & 0 & 0.2 & 0.8 \end{bmatrix} \end{matrix}, \quad \begin{matrix} 1 \leftrightarrow 2, \\ 3 \leftrightarrow 4 \leftrightarrow 5. \end{matrix}$

► $\mathbf{P} = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} 1 \\ 2 \end{matrix} & \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix} \end{matrix}, \quad \pi \mathbf{P} = \pi \Rightarrow \pi = (0.5, 0.5).$

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

$$p(y_{t+1} \mid y_0, y_1, \dots, y_t) = p(y_{t+1} \mid y_t) \quad \cdots \rightarrow \textcircled{y_{t-2}} \rightarrow \textcircled{y_{t-1}} \rightarrow \textcircled{y_t} \rightarrow \textcircled{y_{t+1}} \rightarrow \cdots$$

$$\blacktriangleright \mathbf{P} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0.4 & 0.4 & 0.2 \\ 0 & 0 & 0 & 0.2 & 0.8 \end{bmatrix} \end{matrix}, \quad \begin{matrix} 1 \leftrightarrow 2, \\ 3 \leftrightarrow 4 \leftrightarrow 5. \end{matrix}$$

$$\blacktriangleright \mathbf{P} = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} 1 \\ 2 \end{matrix} & \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix} \end{matrix}, \quad \boldsymbol{\pi} \mathbf{P} = \boldsymbol{\pi} \Rightarrow \boldsymbol{\pi} = (0.5, 0.5).$$

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

$$p(y_{t+1} \mid y_0, y_1, \dots, y_t) = p(y_{t+1} \mid y_t) \quad \cdots \rightarrow \textcircled{y_{t-2}} \rightarrow \textcircled{y_{t-1}} \rightarrow \textcircled{y_t} \rightarrow \textcircled{y_{t+1}} \rightarrow \cdots$$

$$\blacktriangleright \mathbf{P} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0.4 & 0.6 & 0 & 0 & 0 \\ 0.6 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0.7 & 0 \\ 0 & 0 & 0.4 & 0.4 & 0.2 \\ 0 & 0 & 0 & 0.2 & 0.8 \end{bmatrix} \end{matrix}, \quad \begin{matrix} 1 \leftrightarrow 2, \\ 3 \leftrightarrow 4 \leftrightarrow 5. \end{matrix}$$

$$\blacktriangleright \mathbf{P} = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} 1 \\ 2 \end{matrix} & \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix} \end{matrix}, \quad \boldsymbol{\pi} \mathbf{P} = \boldsymbol{\pi} \Rightarrow \boldsymbol{\pi} = (0.5, 0.5).$$

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

$$p(y_{t+1} \mid y_0, y_1, \dots, y_t) = p(y_{t+1} \mid y_t) \quad \cdots \rightarrow y_{t-2} \rightarrow y_{t-1} \rightarrow y_t \rightarrow y_{t+1} \rightarrow \cdots$$

- **Detailed Balance Equation:** $\pi^*(x)p(x, y) = \pi^*(y)p(y, x) \quad \forall (x, y) \in \mathcal{Y}^2$.
- **Reversibility:** A Markov chain is reversible if the detailed balanced equation holds.
- Detailed balance equation implies $\pi^* = \pi^* P$.
 $(\pi^* P)_y = \sum_x \pi^*(x)p(x, y) = \sum_x \pi^*(y)p(y, x) = \pi^*(y) \sum_x p(y, x) = \pi^*(y)$.
- This implies $\pi^* = \pi$, the stationary distribution of P .

$$p(y_{t+1} \mid y_0, y_1, \dots, y_t) = p(y_{t+1} \mid y_t) \quad \cdots \rightarrow \textcircled{y_{t-2}} \rightarrow \textcircled{y_{t-1}} \rightarrow \textcircled{y_t} \rightarrow \textcircled{y_{t+1}} \rightarrow \cdots$$

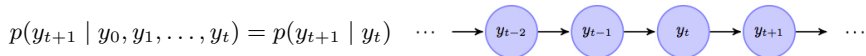
- **Detailed Balance Equation:** $\pi^*(x)p(x, y) = \pi^*(y)p(y, x) \quad \forall (x, y) \in \mathcal{Y}^2$.
- **Reversibility:** A Markov chain is reversible if the detailed balanced equation holds.
- Detailed balance equation implies $\pi^* = \pi^* P$.
 $(\pi^* P)_y = \sum_x \pi^*(x)p(x, y) = \sum_x \pi^*(y)p(y, x) = \pi^*(y) \sum_x p(y, x) = \pi^*(y)$.
- This implies $\pi^* = \pi$, the stationary distribution of P .

$$p(y_{t+1} \mid y_0, y_1, \dots, y_t) = p(y_{t+1} \mid y_t) \quad \cdots \rightarrow y_{t-2} \rightarrow y_{t-1} \rightarrow y_t \rightarrow y_{t+1} \rightarrow \cdots$$

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

$$(\pi^* \mathbf{P})_y = \sum_x \pi^*(x)p(x, y) = \sum_x \pi^*(y)p(y, x) = \pi^*(y) \sum_x p(y, x) = \pi^*(y).$$

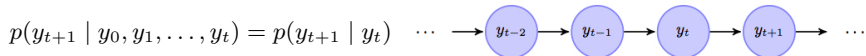
- This implies $\pi^* = \pi$, the stationary distribution of \mathbf{P} .



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

$$(\pi^* \mathbf{P})_y = \sum_x \pi^*(x)p(x, y) = \sum_x \pi^*(y)p(y, x) = \pi^*(y) \sum_x p(y, x) = \pi^*(y).$$

- This implies $\pi^* = \pi$, the stationary distribution of \mathbf{P} .



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

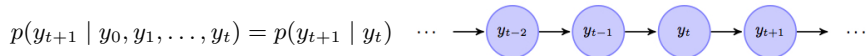
$$p(y_{t+1} \mid y_0, y_1, \dots, y_t) = p(y_{t+1} \mid y_t) \quad \cdots \rightarrow \textcircled{y_{t-2}} \rightarrow \textcircled{y_{t-1}} \rightarrow \textcircled{y_t} \rightarrow \textcircled{y_{t+1}} \rightarrow \cdots$$

- **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:** $\pi(x) p(x, y) = \pi(y) p(y, x) \quad \forall (x, y) \in \mathcal{Y}^2$.
- **Stationary Distribution:** $\pi(y) = \int \pi(x) p(x, y) dx$.

Markov Chains - Continuous Case

$$p(y_{t+1} \mid y_0, y_1, \dots, y_t) = p(y_{t+1} \mid y_t) \quad \cdots \rightarrow \textcircled{y_{t-2}} \rightarrow \textcircled{y_{t-1}} \rightarrow \textcircled{y_t} \rightarrow \textcircled{y_{t+1}} \rightarrow \cdots$$

- **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: $\pi(x) p(x, y) = \pi(y) p(y, x) \quad \forall (x, y) \in \mathcal{Y}^2$.
- Stationary Distribution: $\pi(y) = \int \pi(x) p(x, y) dx$.



- **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:** $\pi(x)p(x, y) = \pi(y)p(y, x) \quad \forall (x, y) \in \mathcal{Y}^2.$
- **Stationary Distribution:** $\pi(y) = \int \pi(x)p(x, y)dx.$

$$p(y_{t+1} \mid y_0, y_1, \dots, y_t) = p(y_{t+1} \mid y_t) \quad \cdots \rightarrow \textcircled{y_{t-2}} \rightarrow \textcircled{y_{t-1}} \rightarrow \textcircled{y_t} \rightarrow \textcircled{y_{t+1}} \rightarrow \cdots$$

- **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:** $\pi(x)p(x, y) = \pi(y)p(y, x) \quad \forall (x, y) \in \mathcal{Y}^2.$
- **Stationary Distribution:** $\pi(y) = \int \pi(x)p(x, y)dx.$

Metropolis-Hastings Sampler

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

Metropolis-Hastings Sampler

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

Metropolis-Hastings Sampler

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

Metropolis-Hastings Sampler

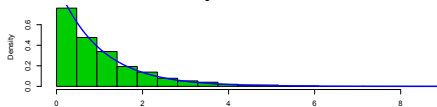
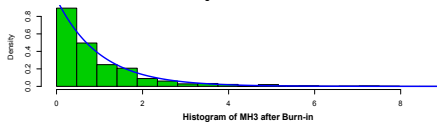
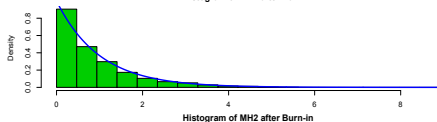
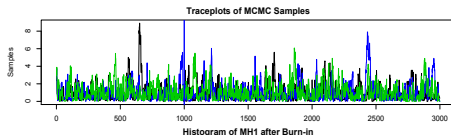
$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

Metropolis-Hastings Sampler - Sampling from an Exponential

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

- ▶ Target: $p(\theta) = \exp(-\theta)$
- ▶ Proposal: $q(\theta^* \mid \theta^{(t-1)}) = \text{Normal}(\theta^* \mid \theta^{(t-1)}, \sigma_\theta^2)$
- ▶ Start with some $\theta^{(0)}$.
- ▶ Propose θ^* according to $q(\theta^* \mid \theta^{(t-1)})$.
- ▶ Compute $\alpha = \min \left\{ 1, \frac{p(\theta^*)}{p(\theta^{(t-1)})} \right\}$.
- ▶ Draw $r = \text{Unif}(0, 1)$.
- ▶ If $r \leq \alpha$, set $\theta^{(t)} = \theta^*$, otherwise, set $\theta^{(t)} = \theta^{(t-1)}$.



Metropolis-Hastings Sampler as a Markov Chain

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

- 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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

- 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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

- 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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

- 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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

► 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: $q(\theta^* \mid \theta^{(t-1)}) = \text{Normal}(\theta^* \mid \theta^{(t-1)}, \sigma_\theta^2)$

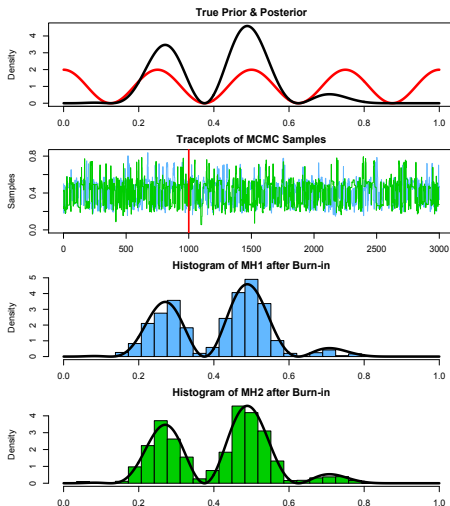
► Start with some $\theta^{(0)}$.

► Propose θ^* according to $q(\theta^* \mid \theta^{(t-1)})$.

► Compute $\alpha = \min \left\{ 1, \frac{p(\theta^* \mid \mathbf{y}_{1:n})}{p(\theta^{(t-1)} \mid \mathbf{y}_{1:n})} \right\}$.

► Draw $r = \text{Unif}(0, 1)$.

► If $r \leq \alpha$, set $\theta^{(t)} = \theta^*$, otherwise, set $\theta^{(t)} = \theta^{(t-1)}$.



Metropolis-Hastings Sampler - Mixing

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \dots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \dots$$

- **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: $p(\theta) \propto \cos^2(4\pi\theta)$

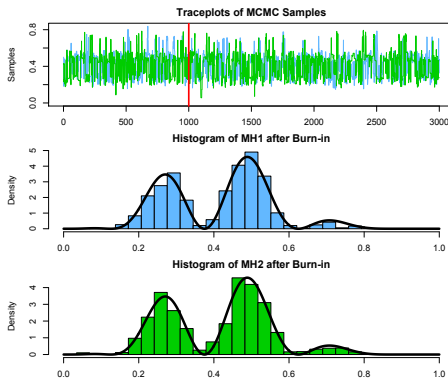
► Target: $p(\theta \mid \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta) \theta^s (1 - \theta)^{n-s}$

► Proposal: $q(\theta^* \mid \theta^{(t-1)}) = \text{Normal}(\theta^* \mid \theta^{(t-1)}, \mathbf{0.25^2})$

► Starting points:

$$\text{MH1: } \theta^{(0)} = 0.50,$$

$$\text{MH2: } \theta^{(0)} = 0.25$$

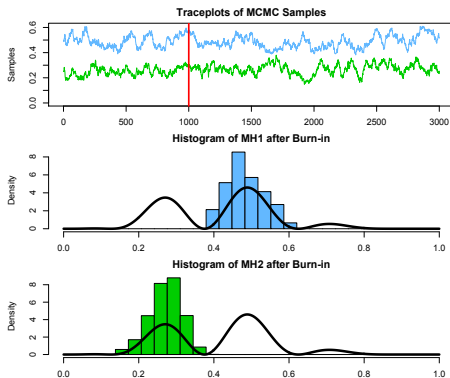


Metropolis-Hastings Sampler - Mixing

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \dots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \dots$$

- **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: $q(\theta^* \mid \theta^{(t-1)}) = \text{Normal}(\theta^* \mid \theta^{(t-1)}, \mathbf{0.01^2})$
- ▶ Starting points:
MH1: $\theta^{(0)} = 0.50$,
MH2: $\theta^{(0)} = 0.25$



Metropolis-Hastings Sampler - Mixing

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

- **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: $p(\theta) \propto \cos^2(4\pi\theta)$

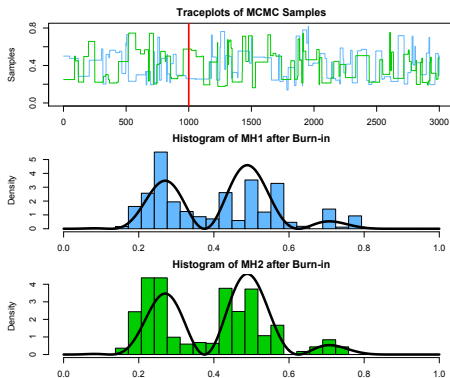
► Target: $p(\theta \mid \mathbf{y}_{1:n}) \propto \cos^2(4\pi\theta) \theta^s (1 - \theta)^{n-s}$

► Proposal: $q(\theta^* \mid \theta^{(t-1)}) = \text{Normal}(\theta^* \mid \theta^{(t-1)}, 4^2)$

► Starting points:

$$\text{MH1: } \theta^{(0)} = 0.50,$$

$$\text{MH2: } \theta^{(0)} = 0.25$$



Metropolis-Hastings Sampler - Simulated Annealing

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

Metropolis-Hastings Sampler:

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

Metropolis-Hastings Sampler with Simulated Annealing:

- Modify the acceptance probability as $\alpha = \min \left\{ 1, \left\{ \frac{p(\theta^*)q(\theta^* \rightarrow \theta^{(t-1)})}{p(\theta^{(t-1)})q(\theta^{(t-1)} \rightarrow \theta^*)} \right\}^{1/T(t)} \right\}$
 - 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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

Metropolis-Hastings Sampler:

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

Metropolis-Hastings Sampler with Simulated Annealing:

- Modify the acceptance probability as $\alpha = \min \left\{ 1, \left\{ \frac{p(\theta^*)q(\theta^* \rightarrow \theta^{(t-1)})}{p(\theta^{(t-1)})q(\theta^{(t-1)} \rightarrow \theta^*)} \right\}^{1/T(t)} \right\}$
 - 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:
$$T(t) = \max \left\{ T_0^{1-t/n}, 1 \right\}.$$

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \longrightarrow \theta_{t-2} \longrightarrow \theta_{t-1} \longrightarrow \theta_t \longrightarrow \theta_{t+1} \longrightarrow \cdots$$

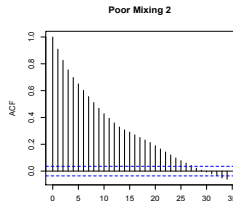
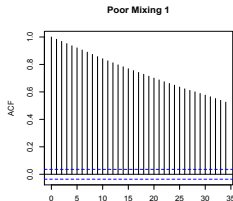
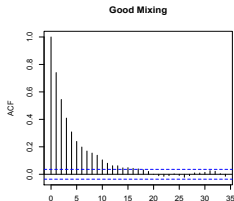
- **Autocorrelation:** The k^{th} order autocorrelation based on n samples after burn-in:

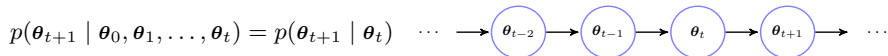
$$\rho_k = \frac{\text{cov}(\theta^{(t)}, \theta^{(t+k)})}{\text{var}(\theta^{(t)})} \hat{=} \frac{\sum_{t=1}^{n-k} (\theta^{(t)} - \bar{\theta})(\theta^{(t+k)} - \bar{\theta})}{\sum_{t=1}^{n-k} (\theta^{(t)} - \bar{\theta})^2} = \hat{\rho}_k, \quad \bar{\theta} = \frac{1}{n} \sum_{t=1}^n \theta^{(t)}.$$

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

$$\rho_k = \frac{\text{cov}(\theta^{(t)}, \theta^{(t+k)})}{\text{var}(\theta^{(t)})} \hat{=} \frac{\sum_{t=1}^{n-k} (\theta^{(t)} - \bar{\theta})(\theta^{(t+k)} - \bar{\theta})}{\sum_{t=1}^{n-k} (\theta^{(t)} - \bar{\theta})^2} = \hat{\rho}_k, \quad \bar{\theta} = \frac{1}{n} \sum_{t=1}^n \theta^{(t)}.$$



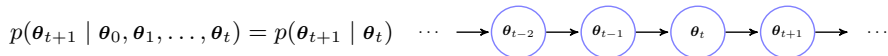


- **Autocorrelation:** The k^{th} order autocorrelation based on n samples after burn-in:

$$\rho_k = \frac{\text{cov}(\theta^{(t)}, \theta^{(t+k)})}{\text{var}(\theta^{(t)})} \hat{=} \frac{\sum_{t=1}^{n-k} (\theta^{(t)} - \bar{\theta})(\theta^{(t+k)} - \bar{\theta})}{\sum_{t=1}^{n-k} (\theta^{(t)} - \bar{\theta})^2} = \hat{\rho}_k, \quad \bar{\theta} = \frac{1}{n} \sum_{t=1}^n \theta^{(t)}.$$

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

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



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

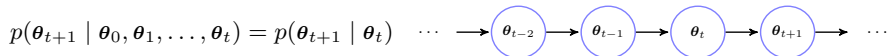
$$\rho_k = \frac{\text{cov}(\theta^{(t)}, \theta^{(t+k)})}{\text{var}(\theta^{(t)})} \hat{=} \frac{\sum_{t=1}^{n-k} (\theta^{(t)} - \bar{\theta})(\theta^{(t+k)} - \bar{\theta})}{\sum_{t=1}^{n-k} (\theta^{(t)} - \bar{\theta})^2} = \hat{\rho}_k, \quad \bar{\theta} = \frac{1}{n} \sum_{t=1}^n \theta^{(t)}.$$

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

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

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

Convergence Diagnostics



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

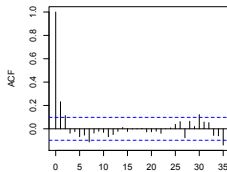
$$\rho_k = \frac{\text{cov}(\theta^{(t)}, \theta^{(t+k)})}{\text{var}(\theta^{(t)})} \hat{=} \frac{\sum_{t=1}^{n-k} (\theta^{(t)} - \bar{\theta})(\theta^{(t+k)} - \bar{\theta})}{\sum_{t=1}^{n-k} (\theta^{(t)} - \bar{\theta})^2} = \hat{\rho}_k, \quad \bar{\theta} = \frac{1}{n} \sum_{t=1}^n \theta^{(t)}.$$

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

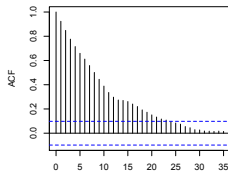
$$\rho_k = \alpha^k, \quad \mathbb{E}(\bar{\theta}_T) = \mu, \quad SE(\bar{\theta}_T) \approx \frac{\sigma}{\sqrt{T}} \frac{1}{(1 - \rho)}.$$

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

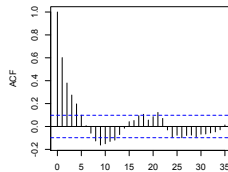
Good Mixing - Thinned



Poor Mixing 1 - Thinned



Poor Mixing 2 - Thinned



$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

$$p(\theta_j) = \int p(\theta_j, \theta_{-j}) d\theta_{-j}$$

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

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

- 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 $\theta^{(0)}$, iteratively sample $\theta^{(t)}$ until convergence.

At each iteration t , for each $j = 1, \dots, p$, sample $\theta_j^{(t)}$ from

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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

$$p(\theta_j) = \int p(\theta_j, \theta_{-j}) d\theta_{-j}$$

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

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

- 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 $\theta^{(0)}$, iteratively sample $\theta^{(t)}$ until convergence.

At each iteration t , for each $j = 1, \dots, p$, sample $\theta_j^{(t)}$ from

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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

$$p(\theta_j) = \int p(\theta_j, \theta_{-j}) d\theta_{-j}$$

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

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

- 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 $\theta^{(0)}$, iteratively sample $\theta^{(t)}$ until convergence.

At each iteration t , for each $j = 1, \dots, p$, sample $\theta_j^{(t)}$ from

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

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

- Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ and $\boldsymbol{\theta}_{-j} = (\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_p)^T$ 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 $j = 1, \dots, p$, sample $\theta_j^{(t)}$ from

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

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

- Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ and $\boldsymbol{\theta}_{-j} = (\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_p)^T$ 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 $j = 1, \dots, p$, sample $\theta_j^{(t)}$ from

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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

$$p(\theta_j) = \int p(\theta_j, \theta_{-j}) d\theta_{-j}$$

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

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

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 $\theta^{(0)}$, iteratively sample $\theta^{(t)}$ until convergence.

At each iteration t , for each $j = 1, \dots, p$, sample $\theta_j^{(t)}$ from

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

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

- Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$ and $\boldsymbol{\theta}_{-j} = (\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_p)^\top$ 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 $j = 1, \dots, p$, sample $\theta_j^{(t)}$ from

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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \longrightarrow \theta_{t-2} \longrightarrow \theta_{t-1} \longrightarrow \theta_t \longrightarrow \theta_{t+1} \longrightarrow \cdots$$

$$\blacktriangleright p(x, y) = \frac{1}{\text{Beta}(\alpha, \beta)} \binom{n}{x} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}, \quad x = 0, 1, \dots, n, \quad y \in (0, 1).$$

- ▶ $p(x \mid y) = \text{Bin}(n, y).$
- ▶ $p(y \mid x) = \text{Beta}(x + \alpha, n - x + \beta).$
- ▶ $n = 10, \quad \alpha = 3, \quad \beta = 3.$

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

$$\blacktriangleright p(x, y) = \frac{1}{\text{Beta}(\alpha, \beta)} \binom{n}{x} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}, \quad x = 0, 1, \dots, n, \quad y \in (0, 1).$$

$$\blacktriangleright p(x \mid y) = \text{Bin}(n, y).$$

$$\blacktriangleright p(y \mid x) = \text{Beta}(x + \alpha, n - x + \beta).$$

$$\blacktriangleright n = 10, \quad \alpha = 3, \quad \beta = 3.$$

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

$$\blacktriangleright p(x, y) = \frac{1}{\text{Beta}(\alpha, \beta)} \binom{n}{x} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}, \quad x = 0, 1, \dots, n, \quad y \in (0, 1).$$

$$\blacktriangleright p(x \mid y) = \text{Bin}(n, y).$$

$$\blacktriangleright p(y \mid x) = \text{Beta}(x + \alpha, n - x + \beta).$$

$$\blacktriangleright n = 10, \quad \alpha = 3, \quad \beta = 3.$$

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

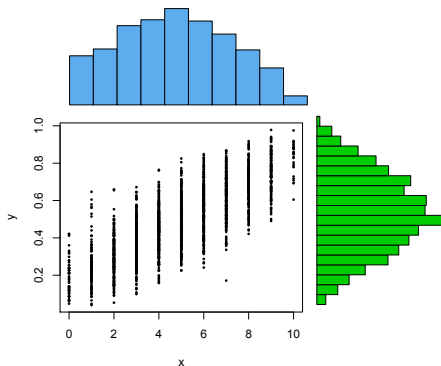
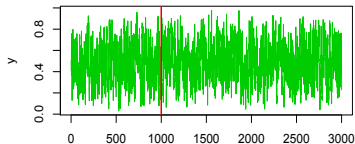
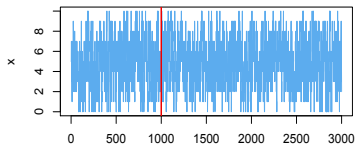
$$\blacktriangleright p(x, y) = \frac{1}{\text{Beta}(\alpha, \beta)} \binom{n}{x} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}, \quad x = 0, 1, \dots, n, \quad y \in (0, 1).$$

- ▶ $p(x \mid y) = \text{Bin}(n, y).$
- ▶ $p(y \mid x) = \text{Beta}(x + \alpha, n - x + \beta).$
- ▶ $n = 10, \quad \alpha = 3, \quad \beta = 3.$

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \dots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \dots$$

► $p(x, y) = \frac{1}{\text{Beta}(\alpha, \beta)} \binom{n}{x} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}, \quad x = 0, 1, \dots, n, \quad y \in (0, 1).$

- $p(x \mid y) = \text{Bin}(n, y).$
- $p(y \mid x) = \text{Beta}(x + \alpha, n - x + \beta).$
- $n = 10, \quad \alpha = 3, \quad \beta = 3.$



Gibbs Sampler as an MH Sampler

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

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

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

- $P_{j,t}^G = q(\theta_j^{(t)} \rightarrow \theta_j^*)$ is the proposal to move from $\theta_j^{(t)}$ to θ_j^* where

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

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

- Also, $\theta_{-j}^* = \theta_{-j}^{(t)}$ since other components do not change.

- Then

$$\begin{aligned} \alpha &= \frac{p(\theta_j^* \mid \mathbf{y}_{1:n})}{p(\theta_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{P_{j,t}^G(\theta_j^{(t)} \mid \theta_j^*)}{P_{j,t}^G(\theta_j^* \mid \theta_j^{(t)})} = \frac{p(\theta_j^* \mid \mathbf{y}_{1:n})}{p(\theta_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta_j^{(t-1)} \mid \theta_{-j}^*, \mathbf{y}_{1:n})}{p(\theta_j^* \mid \theta_{-j}^{(t)}, \mathbf{y}_{1:n})} \\ &= \frac{p(\theta_j^* \mid \theta_{-j}^*, \mathbf{y}_{1:n})}{p(\theta_j^{(t-1)} \mid \theta_{-j}^{(t)}, \mathbf{y}_{1:n})} \times \frac{p(\theta_{-j}^* \mid \mathbf{y}_{1:n})}{p(\theta_{-j}^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta_j^{(t-1)} \mid \theta_{-j}^*, \mathbf{y}_{1:n})}{p(\theta_j^* \mid \theta_{-j}^{(t)}, \mathbf{y}_{1:n})} = 1. \end{aligned}$$

Gibbs Sampler as an MH Sampler

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

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

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

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

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

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

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

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

- Then

$$\begin{aligned} \alpha &= \frac{p(\boldsymbol{\theta}_j^* \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{P_{j,t}^G(\boldsymbol{\theta}_j^{(t)} \mid \boldsymbol{\theta}_j^*)}{P_{j,t}^G(\boldsymbol{\theta}_j^* \mid \boldsymbol{\theta}_j^{(t)})} = \frac{p(\boldsymbol{\theta}_j^* \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\boldsymbol{\theta}_j^{(t-1)} \mid \boldsymbol{\theta}_{-j}^*, \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_j^* \mid \boldsymbol{\theta}_{-j}^{(t)}, \mathbf{y}_{1:n})} \\ &= \frac{p(\boldsymbol{\theta}_j^* \mid \boldsymbol{\theta}_{-j}^*, \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_j^{(t-1)} \mid \boldsymbol{\theta}_{-j}^{(t)}, \mathbf{y}_{1:n})} \times \frac{p(\boldsymbol{\theta}_{-j}^* \mid \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_{-j}^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\boldsymbol{\theta}_j^{(t-1)} \mid \boldsymbol{\theta}_{-j}^*, \mathbf{y}_{1:n})}{p(\boldsymbol{\theta}_j^* \mid \boldsymbol{\theta}_{-j}^{(t)}, \mathbf{y}_{1:n})} = 1. \end{aligned}$$

Gibbs Sampler as an MH Sampler

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

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

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

- $P_{j,t}^G = q(\theta_j^{(t)} \rightarrow \theta_j^*)$ is the proposal to move from $\theta_j^{(t)}$ to θ_j^* where

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

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

- Also, $\theta_{-j}^* = \theta_{-j}^{(t)}$ since other components do not change.

- Then

$$\begin{aligned} \alpha &= \frac{p(\theta_j^* \mid \mathbf{y}_{1:n})}{p(\theta_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{P_{j,t}^G(\theta_j^{(t)} \mid \theta_j^*)}{P_{j,t}^G(\theta_j^* \mid \theta_j^{(t)})} = \frac{p(\theta_j^* \mid \mathbf{y}_{1:n})}{p(\theta_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta_j^{(t-1)} \mid \theta_{-j}^*, \mathbf{y}_{1:n})}{p(\theta_j^* \mid \theta_{-j}^{(t)}, \mathbf{y}_{1:n})} \\ &= \frac{p(\theta_j^* \mid \theta_{-j}^*, \mathbf{y}_{1:n})}{p(\theta_j^{(t-1)} \mid \theta_{-j}^{(t)}, \mathbf{y}_{1:n})} \times \frac{p(\theta_{-j}^* \mid \mathbf{y}_{1:n})}{p(\theta_{-j}^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta_j^{(t-1)} \mid \theta_{-j}^*, \mathbf{y}_{1:n})}{p(\theta_j^* \mid \theta_{-j}^{(t)}, \mathbf{y}_{1:n})} = 1. \end{aligned}$$

Gibbs Sampler as an MH Sampler

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

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

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

- $P_{j,t}^G = q(\theta_j^{(t)} \rightarrow \theta_j^*)$ is the proposal to move from $\theta_j^{(t)}$ to θ_j^* where

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

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

- Also, $\theta_{-j}^* = \theta_{-j}^{(t)}$ since other components do not change.

- Then

$$\begin{aligned} \alpha &= \frac{p(\theta_j^* \mid \mathbf{y}_{1:n})}{p(\theta_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{P_{j,t}^G(\theta_j^{(t)} \mid \theta_j^*)}{P_{j,t}^G(\theta_j^* \mid \theta_j^{(t)})} = \frac{p(\theta_j^* \mid \mathbf{y}_{1:n})}{p(\theta_j^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta_j^{(t-1)} \mid \theta_{-j}^*, \mathbf{y}_{1:n})}{p(\theta_j^* \mid \theta_{-j}^{(t)}, \mathbf{y}_{1:n})} \\ &= \frac{p(\theta_j^* \mid \theta_{-j}^*, \mathbf{y}_{1:n})}{p(\theta_j^{(t-1)} \mid \theta_{-j}^{(t)}, \mathbf{y}_{1:n})} \times \frac{p(\theta_{-j}^* \mid \mathbf{y}_{1:n})}{p(\theta_{-j}^{(t)} \mid \mathbf{y}_{1:n})} \times \frac{p(\theta_j^{(t-1)} \mid \theta_{-j}^*, \mathbf{y}_{1:n})}{p(\theta_j^* \mid \theta_{-j}^{(t)}, \mathbf{y}_{1:n})} = 1. \end{aligned}$$

Normal Model Under Conjugate Prior (From Module V)

$y_1, \dots, y_n \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$ with μ, σ^2 both unknown

► **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp \left[-\frac{1}{2\sigma^2} \left\{ (n-1)s^2 + n(\bar{y} - \mu)^2 \right\} \right]$

► **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2)$

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

► **Normal-Inverse-Gamma Posterior:**

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

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

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

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

► **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)

$y_1, \dots, y_n \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$ with μ, σ^2 both unknown

► **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp \left[-\frac{1}{2\sigma^2} \left\{ (n-1)s^2 + n(\bar{y} - \mu)^2 \right\} \right]$

► **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2)$

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

► **Normal-Inverse-Gamma Posterior:**

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

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

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

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

► **Gibbs: Iteratively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^M$ from the Conditional Posteriors:**

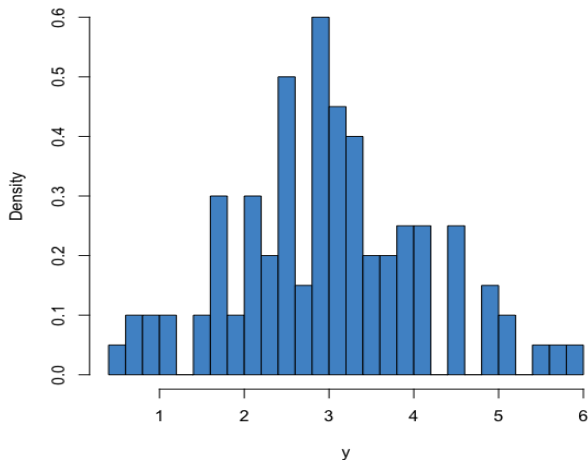
► $p(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \text{Normal}(\mu_n, \sigma^2/\kappa_n)$

► $p(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \text{Inv-Ga}[(\nu_n + 1)/2, \{\kappa_n(\mu - \mu_n)^2 + \nu_n\sigma_n^2\}/2]$

► **Collapsed: Collectively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^M$ from the Marginal Posteriors:**

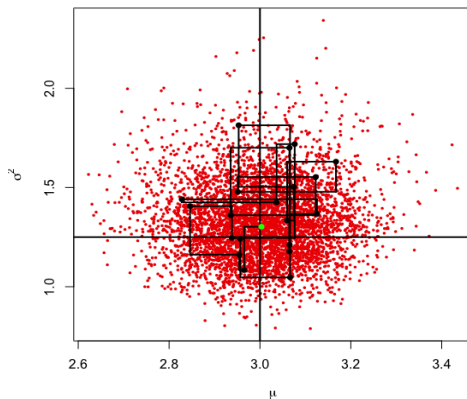
► $p(\mu \mid \mathbf{y}_{1:n}) = t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n)$ ► $p(\sigma^2 \mid \mathbf{y}_{1:n}) = \text{Inv-Ga}(\nu_n/2, \nu_n\sigma_n^2/2)$

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



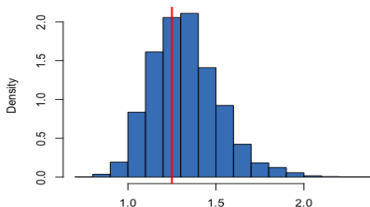
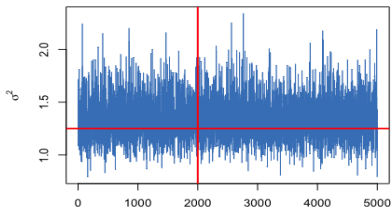
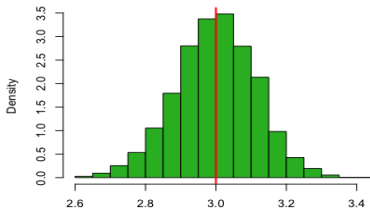
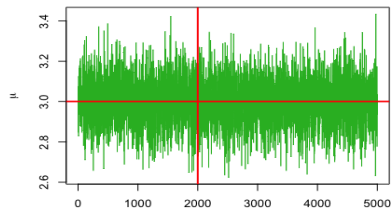
Gibbs Sampler for Normal Model Under NIG Prior - Sample Path

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



Gibbs Sampler for Normal Model Under NIG Prior - Trace Plots

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



$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

Moments:

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

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

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

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

- 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

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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

Moments:

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

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

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

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

- 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

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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

Moments:

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

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

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

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

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

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

$$p(\theta_{t+1} \mid \theta_0, \theta_1, \dots, \theta_t) = p(\theta_{t+1} \mid \theta_t) \quad \cdots \rightarrow \theta_{t-2} \rightarrow \theta_{t-1} \rightarrow \theta_t \rightarrow \theta_{t+1} \rightarrow \cdots$$

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

Moments:

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

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

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

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

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

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

Normal Model Under NIG Prior - Rao-Blackwellization

$y_1, \dots, y_n \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$ with μ, σ^2 both unknown

► **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp \left[-\frac{1}{2\sigma^2} \left\{ (n-1)s^2 + n(\bar{y} - \mu)^2 \right\} \right]$

► **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2)$

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

► **Normal-Inverse-Gamma Posterior:**

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

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

► **Gibbs: Iteratively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^M$ from the Conditional Posteriors:**

$$\begin{aligned} \text{► } p(\mu \mid \sigma^2, \mathbf{y}_{1:n}) &= \text{Normal}(\mu_n, \sigma^2/\kappa_n) & \Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) &\hat{=} \frac{1}{M} \sum_{m=1}^M \mu^{(m)} \\ \text{► } p(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) &= \text{Inv-Ga}[(\nu_n + 1)/2, \{\kappa_n(\mu - \mu_n)^2 + \nu_n\sigma_n^2\}/2] & \Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) &\hat{=} \frac{1}{M} \sum_{m=1}^M \sigma^{2(m)} \end{aligned}$$

► **Collapsed: Collectively Sample from the Marginal Posteriors:**

$$\begin{aligned} \text{► } p(\mu \mid \mathbf{y}_{1:n}) &= t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n) & \Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) &\hat{=} \frac{1}{M} \sum_{m=1}^M \mu^{(m)} \\ \text{► } p(\sigma^2 \mid \mathbf{y}_{1:n}) &= \text{Inv-Ga}(\nu_n/2, \nu_n\sigma_n^2/2) & \Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) &\hat{=} \frac{1}{M} \sum_{m=1}^M \sigma^{2(m)} \end{aligned}$$

► **Rao-Blackwellized: Iteratively Sample from the Conditional Posteriors:**

$$\begin{aligned} \text{► } \mathbb{E}(\mu \mid \sigma^2, \mathbf{y}_{1:n}) &= \mu_{n|\sigma^2} = \mu_n = (\kappa_0\mu_0 + n\bar{y})/(\kappa_0 + n) \\ \text{► } \mathbb{E}(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) &= \sigma_{n|\mu}^2 = \{\kappa_n(\mu - \mu_n)^2 + \nu_n\sigma_n^2\}/(\nu_n - 1) \end{aligned}$$

Normal Model Under NIG Prior - Rao-Blackwellization

$y_1, \dots, y_n \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$ with μ, σ^2 both unknown

► **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp \left[-\frac{1}{2\sigma^2} \left\{ (n-1)s^2 + n(\bar{y} - \mu)^2 \right\} \right]$

► **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2)$

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

► **Normal-Inverse-Gamma Posterior:**

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

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

► **Gibbs: Iteratively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^M$ from the Conditional Posteriors:**

$$\begin{aligned} \text{► } p(\mu \mid \sigma^2, \mathbf{y}_{1:n}) &= \text{Normal}(\mu_n, \sigma^2/\kappa_n) & \Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \mu^{(m)} \\ \text{► } p(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) &= \text{Inv-Ga}[(\nu_n + 1)/2, \{\kappa_n(\mu - \mu_n)^2 + \nu_n\sigma_n^2\}/2] & \Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \sigma^{2(m)} \end{aligned}$$

► **Collapsed: Collectively Sample from the Marginal Posteriors:**

$$\begin{aligned} \text{► } p(\mu \mid \mathbf{y}_{1:n}) &= t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n) & \Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \mu^{(m)} \\ \text{► } p(\sigma^2 \mid \mathbf{y}_{1:n}) &= \text{Inv-Ga}(\nu_n/2, \nu_n\sigma_n^2/2) & \Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \sigma^{2(m)} \end{aligned}$$

► **Rao-Blackwellized: Iteratively Sample from the Conditional Posteriors:**

$$\begin{aligned} \text{► } \mathbb{E}(\mu \mid \sigma^2, \mathbf{y}_{1:n}) &= \mu_{n|\sigma^2} = \mu_n = (\kappa_0\mu_0 + n\bar{y})/(\kappa_0 + n) \\ \text{► } \mathbb{E}(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) &= \sigma_{n|\mu}^2 = \{\kappa_n(\mu - \mu_n)^2 + \nu_n\sigma_n^2\}/(\nu_n - 1) \end{aligned}$$

Normal Model Under NIG Prior - Rao-Blackwellization

$y_1, \dots, y_n \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$ with μ, σ^2 both unknown

► **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp \left[-\frac{1}{2\sigma^2} \left\{ (n-1)s^2 + n(\bar{y} - \mu)^2 \right\} \right]$

► **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2)$

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

► **Normal-Inverse-Gamma Posterior:**

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

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

► **Gibbs: Iteratively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^M$ from the Conditional Posteriors:**

$$\text{► } p(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \text{Normal}(\mu_n, \sigma^2/\kappa_n) \quad \Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \mu^{(m)}$$

$$\text{► } p(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \text{Inv-Ga}[(\nu_n + 1)/2, \{\kappa_n(\mu - \mu_n)^2 + \nu_n\sigma_n^2\}/2] \quad \Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \sigma^{2(m)}$$

► **Collapsed: Collectively Sample from the Marginal Posteriors:**

$$\text{► } p(\mu \mid \mathbf{y}_{1:n}) = t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n) \quad \Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \mu^{(m)}$$

$$\text{► } p(\sigma^2 \mid \mathbf{y}_{1:n}) = \text{Inv-Ga}(\nu_n/2, \nu_n\sigma_n^2/2) \quad \Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \sigma^{2(m)}$$

► **Rao-Blackwellized: Iteratively Sample from the Conditional Posteriors:**

$$\text{► } \mathbb{E}(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \mu_{n|\sigma^2} = \mu_n = (\kappa_0\mu_0 + n\bar{y})/(\kappa_0 + n) \quad \Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \hat{=} \mu_n$$

$$\text{► } \mathbb{E}(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \sigma_{n|\mu}^2 = \{\kappa_n(\mu - \mu_n)^2 + \nu_n\sigma_n^2\}/(\nu_n - 1) \quad \Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \sigma_{n|\mu}^{2(m)}$$

Normal Model Under NIG Prior - Rao-Blackwellization

$y_1, \dots, y_n \stackrel{iid}{\sim} \text{Normal}(\mu, \sigma^2)$ with μ, σ^2 both unknown

► **Normal Likelihood:** $p(\mathbf{y}_{1:n} \mid \mu, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp \left[-\frac{1}{2\sigma^2} \left\{ (n-1)s^2 + n(\bar{y} - \mu)^2 \right\} \right]$

► **Normal-Inverse-Gamma Prior:** $(\mu, \sigma^2) \sim \text{NIG}(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2)$

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

► **Normal-Inverse-Gamma Posterior:**

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

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

► **Gibbs: Iteratively Sample $\{(\mu^{(m)}, \sigma^{2(m)})\}_{m=1}^M$ from the Conditional Posteriors:**

$$\text{► } p(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \text{Normal}(\mu_n, \sigma^2/\kappa_n) \quad \Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \mu^{(m)}$$

$$\text{► } p(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \text{Inv-Ga}[(\nu_n + 1)/2, \{\kappa_n(\mu - \mu_n)^2 + \nu_n\sigma_n^2\}/2] \quad \Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \sigma^{2(m)}$$

► **Collapsed: Collectively Sample from the Marginal Posteriors:**

$$\text{► } p(\mu \mid \mathbf{y}_{1:n}) = t_{\nu_n}(\mu_n, \sigma_n^2/\kappa_n) \quad \Rightarrow \mathbb{E}(\mu \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \mu^{(m)}$$

$$\text{► } p(\sigma^2 \mid \mathbf{y}_{1:n}) = \text{Inv-Ga}(\nu_n/2, \nu_n\sigma_n^2/2) \quad \Rightarrow \mathbb{E}(\sigma^2 \mid \mathbf{y}_{1:n}) \hat{=} \frac{1}{M} \sum_{m=1}^M \sigma^{2(m)}$$

► **Rao-Blackwellized: Iteratively Sample from the Conditional Posteriors:**

$$\text{► } \mathbb{E}(\mu \mid \sigma^2, \mathbf{y}_{1:n}) = \mu_n \mid \sigma^2 = \mu_n = (\kappa_0\mu_0 + n\bar{y})/(\kappa_0 + n) \quad \rightarrow \text{Does not depend on the samples at all!}$$

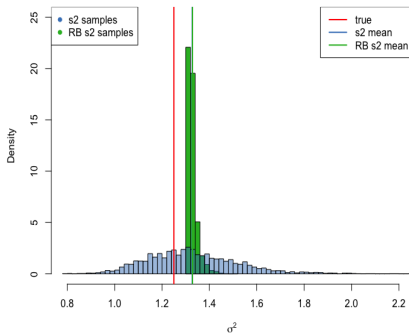
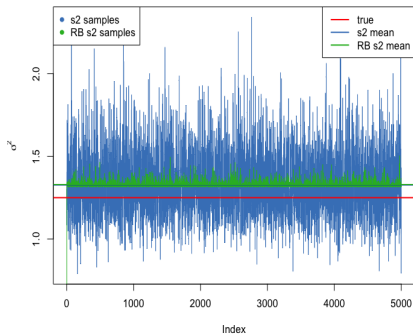
$$\text{► } \mathbb{E}(\sigma^2 \mid \mu, \mathbf{y}_{1:n}) = \sigma_{n|\mu}^2 = \{\kappa_n(\mu - \mu_n)^2 + \nu_n\sigma_n^2\}/(\nu_n - 1) \rightarrow \text{Depends only on the samples of } \mu!$$

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

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

$$\hat{\sigma}_{\text{post-mean}}^2 = \frac{1}{M} \sum_{m=1}^M \sigma^{2(m)}$$

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



- Monte Carlo integration

- Approximates integrals writing it as an expectation then sampling from the corresponding density
- Importance sampling

• Approximates integrals by sampling from a target density p and computing the expectation of a function f with respect to p . This is done by sampling from a reference density q in order to sample from p .

- Sampling from a target density when direct sampling is difficult

• Importance sampling

• Markov Chain Monte Carlo

• Metropolis-Hastings algorithm

- Monte Carlo integration
 - Approximates integrals writing it as an expectation then sampling from the corresponding density
 - Importance sampling
 - Allows some cases important when sampling from a target density is difficult
 - Draw weighted samples from a reference density that is easy to sample
 - Weighted sum what the target density becomes with respect to the reference density
 - Sampling from a target density when direct sampling is difficult

- 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

- 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

- 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

- 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

- 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
 - Resamples draws from a reference density with importance sampling weights
 - Markov chain Monte Carlo
 - Generates draws from a target density by iteratively sampling from a proposal density and accepting or rejecting the proposal

- 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

- 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

- 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

- 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

- Ergodicity: the chain converges to a unique stationary distribution
 - Aperiodicity: no cycles of fixed length between states
 - Irreducibility: the chain can move from any state to any other state
 - Recurrence: the chain will visit any state infinitely often

- Metropolis-Hastings sampler

- The Metropolis-Hastings sampler is a general-purpose MCMC algorithm that can be used to sample from any target distribution
 - It is based on the idea of proposing a new state and then accepting or rejecting it based on the ratio of the target distribution to the proposal distribution
 - The acceptance probability is given by $\min(1, \frac{p(z_t)}{p(z_{t-1})} \frac{q(z_{t-1} \mid z_t)}{q(z_t \mid z_{t-1})})$
 - The proposal distribution $q(z_t \mid z_{t-1})$ is chosen to be symmetric, i.e. $q(z_t \mid z_{t-1}) = q(z_{t-1} \mid z_t)$

- Gibbs sampler

- The Gibbs sampler is a special case of the Metropolis-Hastings sampler where the proposal distribution is chosen to be the conditional distribution of the target distribution
 - It is based on the idea of sampling each component of the state vector from its conditional distribution given the other components
 - The Gibbs sampler is often used for sampling from multivariate distributions

- Convergence diagnostics

- Trace plots: plots of the state vector over time
 - Autocorrelation plots: plots of the autocorrelation function

- 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

- Ergodicity: the chain converges to the target distribution
 - Aperiodicity: no cycles of fixed length between states
 - Irreducibility: probability of reaching a state is > 0 in a finite number of steps
 - Recurrence: probability of returning to state i is 1, for all states i

- Metropolis-Hastings sampler

- Generates samples from a target distribution that is difficult to directly sample from
 - Can sample from a target distribution by using a distribution $q(z_t \mid z_{t-1})$ to propose a new state z_t and then accepting or rejecting the proposal according to the target distribution
 - Acceptance ratio depends on the ratio of the target distribution
 - The sampler converges to the target distribution

- Gibbs sampler

- A special case of the Metropolis-Hastings sampler
 - Generates samples from a target distribution by sampling each variable in turn
 - Samples converge to the target distribution

- Convergence diagnostics

- Trace plots and autocorrelation plots
 - Gelman-Rubin statistic

- 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
 - Chooses a proposal distribution $q(z_t \mid z_{t-1})$ and the target distribution $p(z_t)$ of the Markov chain
 - Chooses a random variable z_t from the proposal distribution with some probability
 - Accepts or rejects the variable based on the ratio of $p(z_t)$ and $q(z_t \mid z_{t-1})$
 - The resulting MC samples from a stationary, multi-dimensional distribution
 - Gibbs sampler
 - A special case of the MCMC that always accepts
 - Chooses the sampling distribution $q(z_t \mid z_{t-1})$ as a conditional distribution
 - Samples sequentially from conditional distributions of each variable given all others
 - Convergence diagnostics
 - Burn-in and thinning
 - Trace plots

- 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
 - Sampling from a target distribution that is difficult to draw samples from
 - The target distribution is the target distribution of the Markov chain
 - The proposal distribution is a distribution that is easy to sample from
 - The acceptance probability is a function of the target and proposal distributions
 - The sampler converges to the target distribution if the chain is irreducible and aperiodic
 - Gibbs sampler
 - A special case of the MCMC algorithm that always accepts
 - The target distribution is the target distribution of the Markov chain
 - The sampler converges to the target distribution if the chain is irreducible and aperiodic
 - Convergence diagnostics
 - Trace plots
 - Autocorrelation function
 - Gelman-Rubin diagnostic

- 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
 - The Metropolis-Hastings sampler is a Markov chain Monte Carlo (MCMC) algorithm that generates a sequence of samples from a target distribution p^* by iteratively proposing new states and accepting or rejecting them based on a certain criterion.
 - The proposal distribution $q(z_t \mid z_{t-1})$ is used to generate candidate states z_t from the current state z_{t-1} .
 - The acceptance probability $\alpha(z_{t-1}, z_t)$ is calculated as $\min\left(1, \frac{p^*(z_t)q(z_{t-1} \mid z_t)}{p^*(z_{t-1})q(z_t \mid z_{t-1})}\right)$.
 - If the candidate state is accepted, it becomes the next state in the chain; otherwise, the current state is repeated.
 - The resulting sequence of states forms a Markov chain that converges to the target distribution p^* under certain conditions.
 - Gibbs sampler
 - The Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm that generates a sequence of samples from a target distribution p^* by iteratively sampling each variable from its conditional distribution given the current values of the other variables.
 - The target distribution p^* is assumed to be a joint distribution over a set of variables $z = (z_1, z_2, \dots, z_K)$.
 - The Gibbs sampler iterates over the variables z_1, z_2, \dots, z_K , sampling each z_i from its conditional distribution $p(z_i \mid z_{-i})$, where z_{-i} represents the current values of all other variables.
 - The resulting sequence of states forms a Markov chain that converges to the target distribution p^* under certain conditions.
 - Convergence diagnostics
 - The Gelman-Rubin diagnostic (GR) is a measure of convergence for MCMC chains. It compares the variance of the means of multiple chains to the variance within each chain.
 - The trace plot shows the values of a parameter of interest over the iterations of the MCMC chain.
 - The autocorrelation function (ACF) plot shows the correlation between samples at different lags.
 - The burn-in period is the initial part of the MCMC chain that is discarded to ensure that the samples are from the target distribution.

- 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
 - A Markov chain Monte Carlo (MCMC) algorithm for sampling from a target distribution p^* using a proposal distribution q .
 - The algorithm generates a sequence of states z_1, z_2, \dots, z_T such that the empirical distribution converges to p^* .
 - The acceptance probability is given by $\alpha = \min(1, \frac{p^*(z')q(z \mid z')}{p^*(z)q(z' \mid z)})$.
 - Gibbs sampler
 - A Markov chain Monte Carlo (MCMC) algorithm for sampling from a target distribution p^* using a full conditional distribution q .
 - The algorithm generates a sequence of states z_1, z_2, \dots, z_T such that the empirical distribution converges to p^* .
 - The acceptance probability is given by $\alpha = \min(1, \frac{p^*(z')q(z \mid z')}{p^*(z)q(z' \mid z)})$.
 - 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

- 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 case of the MCMC algorithm that always proposes
 - Each new value is sampled from the conditional distribution of the variable of interest, given the current values of all other variables
 - Converges faster than MCMC when the proposal of each variable is good
 - 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
 - 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
 - Samples from each variable in turn, using the current values of the other variables
 - Requires the conditional distributions to be tractable
 - Convergence is often slower than for Metropolis-Hastings
 - 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
 - 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
 - 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, samples each variable in turn from its conditional distribution given the current values of all other variables
 - 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
 - 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
 - Similar to Metropolis-Hastings, but instead of proposing a new value, each variable is sampled from its conditional distribution
 - 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
 - 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
 - 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
 - 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
 - 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

- 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

- 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

- 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

- 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

- 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
 - 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
 - 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](#)