Intro to Bayesian Computing

Krzysztof Latuszynski
(University of Warwick, UK)

OxWaSP - module 1
The Bayesian setting

- Prior-posterior
- Uncertainty quantification
- MAP and Bayesian estimators

Sampling Probability Distributions 1 - direct approaches

- CLT for Monte Carlo
- Inverse cdf method
- Rejection Sampling
- Importance Sampling
- Sequential Importance Sampling

Sampling Probability distributions 2 - Markov chains

- MCMC
- CLT for MCMC
- Detailed balance
- Metropolis-Hastings
- Gibbs samplers
Prior-Posterior

- let $\theta \in \Theta$ be a parameter of a statistical model, say $M(\theta)$. E.g. $\Theta \in \mathbb{R}^d$, $\Theta \in \mathbb{N}^d$, $\Theta \in \{0, 1\}^d$

- In Bayesian Statistics one assumes $\theta$ is random, i.e. there exists a prior probability distribution $p(\theta)$ on $\Theta$ s.t. in absence of additional information $\theta \sim p(\theta)$.

- $y_1, \ldots, y_n \in \mathbb{Y}^n$ - data

- $l(\theta|y_1, \ldots, y_n)$ - the likelihood function for the model $M(\theta)$

- Example: Consider a diffusion model $M(\theta)$ where $\theta = (\mu, \sigma)$

\[
dX_t = \mu dt + \sigma dB_t
\]

observed at discrete time points $(t_0, t_1, \ldots, t_N)$ as $(x_{t_0}, x_{t_1}, \ldots, x_{t_N})$

- The likelihood function is

\[
l(\theta|x_{t_0}, x_{t_1}, \ldots, x_{t_N}) = \prod_{i=1}^{N} l(\theta|x_{t_i}, x_{t_{i-1}}) = \prod_{i=1}^{N} \phi_N(\mu(t_i-t_{i-1}), \sigma^2(t_i-t_{i-1}))(x_{t_i} - x_{t_{i-1}}).
\]
Prior-Posterior

- Let $\theta \in \Theta$ be a parameter of a statistical model, say $M(\theta)$. 
  E.g. $\Theta \in \mathbb{R}^d$, $\Theta \in \mathbb{N}^d$, $\Theta \in \{0,1\}^d$
- In Bayesian Statistics one assumes $\theta$ is random, i.e. there exists a prior probability distribution $p(\theta)$ on $\Theta$ s.t. in absence of additional information $\theta \sim p(\theta)$.
- $y_1, \ldots, y_n \in \mathbb{Y}^n$ - data
- $l(\theta|y_1, \ldots, y_n)$ - the likelihood function for the model $M(\theta)$
- Example: Consider a diffusion model $M(\theta)$ where $\theta = (\mu, \sigma)$

$$dX_t = \mu dt + \sigma dB_t$$

observed at discrete time points $(t_0, t_1, \ldots, t_N)$ as $(x_{t_0}, x_{t_1}, \ldots, x_{t_N})$

- The likelihood function is

$$l(\theta|x_{t_0}, x_{t_1}, \ldots, x_{t_N}) = \prod_{i=1}^{N} l(\theta|x_{t_i}, x_{t_{i-1}}) = \prod_{i=1}^{N} \phi_N(\mu(t_{i}-t_{i-1}), \sigma^2(t_{i}-t_{i-1}))(x_{t_i} - x_{t_{i-1}}).$$
Prior-Posterior

- let $\theta \in \Theta$ be a parameter of a statistical model, say $M(\theta)$. 
  E.g. $\Theta \in \mathbb{R}^d$, $\Theta \in \mathbb{N}^d$, $\Theta \in \{0, 1\}^d$

- In Bayesian Statistics one assumes $\theta$ is random, i.e. there exists a prior probability distribution $p(\theta)$ on $\Theta$ s.t. in absence of additional information $\theta \sim p(\theta)$.

- $y_1, \ldots, y_n \in \mathbb{Y}^n$ - data

- $l(\theta|y_1, \ldots, y_n)$ - the likelihood function for the model $M(\theta)$

- Example: Consider a diffusion model $M(\theta)$ where $\theta = (\mu, \sigma)$

$$dX_t = \mu dt + \sigma dB_t$$

observed at discrete time points $(t_0, t_1, \ldots, t_N)$ as $(x_{t_0}, x_{t_1}, \ldots, x_{t_N})$

- The likelihood function is

$$l(\theta|x_{t_0}, x_{t_1}, \ldots, x_{t_N}) = \prod_{i=1}^{N} l(\theta|x_{t_i}, x_{t_{i-1}}) = \prod_{i=1}^{N} \phi_N(\mu(t_i-t_{i-1}), \sigma^2(t_i-t_{i-1}))(x_{t_i} - x_{t_{i-1}}).$$
Prior-Posterior

- Let $\theta \in \Theta$ be a parameter of a statistical model, say $M(\theta)$. E.g. $\Theta \in \mathbb{R}^d$, $\Theta \in \mathbb{N}^d$, $\Theta \in \{0, 1\}^d$

- In Bayesian Statistics one assumes $\theta$ is random, i.e. there exists a prior probability distribution $p(\theta)$ on $\Theta$ s.t. in absence of additional information $\theta \sim p(\theta)$.

- $y_1, \ldots, y_n \in \mathbb{Y}^n$ - data

- $l(\theta|y_1, \ldots, y_n)$ - the likelihood function for the model $M(\theta)$

- Example: Consider a diffusion model $M(\theta)$ where $\theta = (\mu, \sigma)$

$$dX_t = \mu dt + \sigma dB_t$$

observed at discrete time points $(t_0, t_1, \ldots, t_N)$ as $(x_{t_0}, x_{t_1}, \ldots, x_{t_N})$

- The likelihood function is

$$l(\theta|x_{t_0}, x_{t_1}, \ldots, x_{t_N}) = \prod_{i=1}^{N} l(\theta|x_{t_i}, x_{t_{i-1}}) = \prod_{i=1}^{N} \phi_N(\mu(t_i-t_{i-1}), \sigma^2(t_i-t_{i-1}))(x_{t_i} - x_{t_{i-1}}).$$
Prior-Posterior

- let $\theta \in \Theta$ be a parameter of a statistical model, say $M(\theta)$. E.g. $\Theta \in \mathbb{R}^d$, $\Theta \in \mathbb{N}^d$, $\Theta \in \{0, 1\}^d$
- In Bayesian Statistics one assumes $\theta$ is random, i.e. there exists a prior probability distribution $p(\theta)$ on $\Theta$ s.t. in absence of additional information $\theta \sim p(\theta)$.
- $y_1, \ldots, y_n \in Y^n$ - data
- $l(\theta|y_1, \ldots, y_n)$ - the likelihood function for the model $M(\theta)$
- Example: Consider a diffusion model $M(\theta)$ where $\theta = (\mu, \sigma)$

$$dX_t = \mu dt + \sigma dB_t$$

observed at discrete time points $(t_0, t_1, \ldots, t_N)$ as $(x_{t_0}, x_{t_1}, \ldots, x_{t_N})$

The likelihood function is

$$l(\theta|x_{t_0}, x_{t_1}, \ldots, x_{t_N}) = \prod_{i=1}^{N} l(\theta|x_{t_i}, x_{t_{i-1}}) = \prod_{i=1}^{N} \phi_N(\mu(t_i-t_{i-1}), \sigma^2(t_i-t_{i-1}))(x_{t_i} - x_{t_{i-1}}).$$
Prior-Posterior

- Let $\theta \in \Theta$ be a parameter of a statistical model, say $M(\theta)$.
  
  E.g. $\Theta \in \mathbb{R}^d$, $\Theta \in \mathbb{N}^d$, $\Theta \in \{0, 1\}^d$

- In Bayesian Statistics one assumes $\theta$ is random, i.e. there exists a prior probability distribution $p(\theta)$ on $\Theta$ s.t. in absence of additional information $\theta \sim p(\theta)$.

- $y_1, \ldots, y_n \in \mathbb{Y}^n$ - data

- $l(\theta|y_1, \ldots, y_n)$ - the likelihood function for the model $M(\theta)$

- Example: Consider a diffusion model $M(\theta)$ where $\theta = (\mu, \sigma)$

\[
dX_t = \mu dt + \sigma dB_t
\]

observed at discrete time points $(t_0, t_1, \ldots, t_N)$ as $(x_{t_0}, x_{t_1}, \ldots, x_{t_N})$

- The likelihood function is

\[
l(\theta|x_{t_0}, x_{t_1}, \ldots, x_{t_N}) = \prod_{i=1}^{N} l(\theta|x_{t_i}, x_{t_{i-1}}) = \prod_{i=1}^{N} \phi_N(\mu(t_i-t_{i-1}), \sigma^2(t_i-t_{i-1}))(x_{t_i} - x_{t_{i-1}}).
\]
The posterior distribution is then

\[
\pi(\theta) = \pi(\theta|y_1, \ldots, y_n) = \frac{p(\theta)l(\theta|y_1, \ldots, y_n)}{\int_{\Theta} p(\theta)l(\theta|y_1, \ldots, y_n)d\theta}.
\]

This posterior summarises uncertainty about the parameter \( \theta \in \Theta \) and is used for all inferential questions like credible sets, decision making, prediction, model choice, etc.

In the diffusion example predicting the value of the diffusion at time \( t > t_N \) would amount to repeating the following steps:

1. sample \( \theta = (\mu, \sigma) \sim \pi(\theta) \)
2. sample \( X_t \sim N(x_{t_N} + \mu(t - t_N), \sigma^2(t - t_N)) \)
The posterior distribution is then

$$
\pi(\theta) = \pi(\theta|y_1, \ldots, y_n) = \frac{p(\theta)l(\theta|y_1, \ldots, y_n)}{\int_{\Theta} p(\theta)l(\theta|y_1, \ldots, y_n) d\theta}.
$$

This posterior summarises uncertainty about the parameter $\theta \in \Theta$ and is used for all inferential questions like credible sets, decision making, prediction, model choice, etc.

In the diffusion example predicting the value of the diffusion at time $t > t_N$ would amount to repeating the following steps:

1. sample $\theta = (\mu, \sigma) \sim \pi(\theta)
2. sample $X_t \sim N(x_{t_N} + \mu(t - t_N), \sigma^2(t - t_N))$
The posterior distribution is then

$$\pi(\theta) = \pi(\theta|y_1, \ldots, y_n) = \frac{p(\theta)l(\theta|y_1, \ldots, y_n)}{\int_{\Theta} p(\theta)l(\theta|y_1, \ldots, y_n) d\theta}.$$ 

This posterior summarises uncertainty about the parameter $\theta \in \Theta$ and is used for all inferential questions like credible sets, decision making, prediction, model choice, etc.

In the diffusion example predicting the value of the diffusion at time $t > t_N$ would amount to repeating the following steps:

1. sample $\theta = (\mu, \sigma) \sim \pi(\theta)$
2. sample $X_t \sim N(x_{t_N} + \mu(t - t_N), \sigma^2(t - t_N))$
One of the classical estimation tasks is to compute the Maximum a Posteriori Estimator (MAP), say $\theta_{MAP}$.

\[
\theta_{MAP} := \arg\max_\theta \pi(\theta) = \arg\max_\theta \left\{ p(\theta) l(\theta|y_1, \ldots, y_n) \right\}
\]

Computing $\theta_{MAP}$ may be nontrivial, especially if $\pi(\theta)$ is multimodal.

There are specialised algorithms for doing this.

Some non-bayesian statistical inference approaches can be rewritten as bayesian MAP estimators.
One of the classical estimation tasks is to compute the Maximum a Posteriori Estimator (MAP), say $\theta_{MAP}$.

$$\theta_{MAP} := \arg\max_\theta \pi(\theta) = \arg\max_\theta \left\{ p(\theta) l(\theta|y_1, \ldots, y_n) \right\}$$

Computing $\theta_{MAP}$ may be nontrivial, especially if $\pi(\theta)$ is multimodal.

There are specialised algorithms for doing this.

Some non-bayesian statistical inference approaches can be rewritten as bayesian MAP estimators.
One of the classical estimation tasks is to compute the Maximum a Posteriori Estimator (MAP), say $\theta_{MAP}$.

$$
\theta_{MAP} := \arg\max_\theta \pi(\theta) = \arg\max_\theta \left\{ p(\theta) l(\theta | y_1, \ldots, y_n) \right\}
$$

Computing $\theta_{MAP}$ may be nontrivial, especially if $\pi(\theta)$ is multimodal.

There are specialised algorithms for doing this.

Some non-bayesian statistical inference approaches can be rewritten as bayesian MAP estimators.
One of the classical estimation tasks is to compute the Maximum a Posteriori Estimator (MAP), say $\theta_{MAP}$.

$$
\theta_{MAP} := \arg\max_{\theta} \pi(\theta) = \arg\max_{\theta} \left\{ p(\theta) l(\theta | y_1, \ldots, y_n) \right\}
$$

Computing $\theta_{MAP}$ may be nontrivial, especially if $\pi(\theta)$ is multimodal.

There are specialised algorithms for doing this.

Some non-bayesian statistical inference approaches can be rewritten as bayesian MAP estimators.
One of the classical estimation tasks is to compute the Maximum a Posteriori Estimator (MAP), say $\theta_{MAP}$.

$$\theta_{MAP} := \arg\max_{\theta} \pi(\theta) = \arg\max_{\theta} \left\{ p(\theta)l(\theta|y_1, \ldots, y_n) \right\}$$

Computing $\theta_{MAP}$ may be nontrivial, especially if $\pi(\theta)$ is multimodal.

There are specialised algorithms for doing this.

Some non-bayesian statistical inference approaches can be rewritten as bayesian MAP estimators.
The Bayesian setting
Sampling Probability Distributions 1 - direct approaches
Sampling Probability distributions 2 - Markov chains

Prior-posterior
Uncertainty quantification
MAP and Bayesian estimators

the Bayesian estimator

- Bayesian estimator is an estimator that minimizes the posterior expected value of a loss function.
- The loss function
  \[ L(\cdot, \cdot) : \Theta \times \Theta \to \mathbb{R} \]
- After seeing data \((y_1, \ldots, y_n)\) we choose an estimator \(\hat{\theta}(y_1, \ldots, y_n)\)
- Its expected loss is
  \[
  \mathbb{E}L(\theta, \hat{\theta}(y_1, \ldots, y_n)) = \int_{Y^n \times \Theta} L(\theta, \hat{\theta}(y_1, \ldots, y_n))m(y_1, \ldots, y_n|\theta)p(\theta) \]
  \[= \int_{Y^n \times \Theta} L(\theta, \hat{\theta}(y_1, \ldots, y_n))\pi(\theta)p(dx) \]
- \(\hat{\theta}(y_1, \ldots, y_n)\) is a Bayesian estimator if it minimizes the above expected loss.
The Bayesian estimator

- Bayesian estimator is an estimator that minimizes the posterior expected value of a loss function.
- The loss function
  \[ L(\cdot, \cdot) : \Theta \times \Theta \to \mathbb{R} \]
- After seeing data \((y_1, \ldots, y_n)\) we choose an estimator \(\hat{\theta}(y_1, \ldots, y_n)\)
- Its expected loss is
  \[
  \mathbb{E}L(\theta, \hat{\theta}(y_1, \ldots, y_n)) = \int_{Y^n \times \Theta} L(\theta, \hat{\theta}(y_1, \ldots, y_n))m(y_1, \ldots, y_n | \theta)p(\theta) \\
  = \int_{Y^n \times \Theta} L(\theta, \hat{\theta}(y_1, \ldots, y_n))\pi(\theta)p(dx)
  \]
- \(\hat{\theta}(y_1, \ldots, y_n)\) is a Bayesian estimator if it minimizes the above expected loss.
the Bayesian estimator

- Bayesian estimator is an estimator that minimizes the posterior expected value of a loss function.
- The loss function
  \[ L(\cdot, \cdot) : \Theta \times \Theta \to \mathbb{R} \]
- After seeing data \((y_1, \ldots, y_n)\) we choose an estimator \(\hat{\theta}(y_1, \ldots, y_n)\)
- Its expected loss is
  \[
  \mathbb{E}L(\theta, \hat{\theta}(y_1, \ldots, y_n)) = \int_{\Theta \times \Theta} L(\theta, \hat{\theta}(y_1, \ldots, y_n))m(y_1, \ldots, y_n | \theta)p(\theta) \]
  \[
  = \int_{\Theta \times \Theta} L(\theta, \hat{\theta}(y_1, \ldots, y_n))\pi(\theta)p(\theta)dx \]
- \(\hat{\theta}(y_1, \ldots, y_n)\) is a Bayesian estimator if it minimizes the above expected loss.
Bayesian estimator is an estimator that minimizes the posterior expected value of a loss function.

The loss function

\[ L(\cdot, \cdot) : \Theta \times \Theta \rightarrow \mathbb{R} \]

After seeing data \((y_1, \ldots, y_n)\) we choose an estimator \(\hat{\theta}(y_1, \ldots, y_n)\)

Its expected loss is

\[
\mathbb{E}L(\theta, \hat{\theta}(y_1, \ldots, y_n)) = \int_{\mathbb{Y}^n \times \Theta} L(\theta, \hat{\theta}(y_1, \ldots, y_n))m(y_1, \ldots, y_n | \theta)p(\theta)
\]

\[= \int_{\mathbb{Y}^n \times \Theta} L(\theta, \hat{\theta}(y_1, \ldots, y_n))\pi(\theta)p(dx)\]

\(\hat{\theta}(y_1, \ldots, y_n)\) is a Bayesian estimator if it minimizes the above expected loss.
The Bayesian estimator

- Bayesian estimator is an estimator that minimizes the posterior expected value of a loss function.
- The loss function
  \[ L(\cdot, \cdot) : \Theta \times \Theta \to \mathbb{R} \]
- After seeing data \((y_1, \ldots, y_n)\) we choose an estimator \(\hat{\theta}(y_1, \ldots, y_n)\)
- Its expected loss is
  \[
  \mathbb{E}L(\theta, \hat{\theta}(y_1, \ldots, y_n)) = \int_{Y^n \times \Theta} L(\theta, \hat{\theta}(y_1, \ldots, y_n))m(y_1, \ldots, y_n | \theta)p(\theta)
  \]
  \[
  = \int_{Y^n \times \Theta} L(\theta, \hat{\theta}(y_1, \ldots, y_n))\pi(\theta)p(dx)
  \]
- \(\hat{\theta}(y_1, \ldots, y_n)\) is a Bayesian estimator if it minimizes the above expected loss.
We consider only the most common choice of quadratic loss function

\[ L(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2 \]

in which case

\[ \hat{\theta}(y_1, \ldots, y_n) = \mathbb{E}_{\pi} \theta \]

so it is the posterior mean.

So computing the Bayesian estimator is computing the integral wrt the posterior

\[ \int_{\Theta} \theta \pi(\theta) \]

Similarly answering other inferential questions like credible sets, posterior variance etc involve computing integrals of the form

\[ \int_{\Theta} f(\theta) \pi(\theta). \]
We consider only the most common choice of quadratic loss function

\[ L(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2 \]

in which case

\[ \hat{\theta}(y_1, \ldots, y_n) = \mathbb{E}_\pi \theta \]

so it is the posterior mean.

So computing the Bayesian estimator is computing the integral wrt the posterior

\[ \int_\Theta \theta \pi(\theta) \]

Similarly answering other inferential questions like credible sets, posterior variance etc involve computing integrals of the form

\[ \int_\Theta f(\theta) \pi(\theta). \]
the Bayesian estimator and computing integrals

- We consider only the most common choice of quadratic loss function
  \[ L(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2 \]

- in which case
  \[ \hat{\theta}(y_1, \ldots, y_n) = \mathbb{E}_\pi \theta \]
  so it is the posterior mean.

- So computing the Bayesian estimator is computing the integral wrt the posterior
  \[ \int_{\Theta} \theta \pi(\theta) \]

- Similarly answering other inferential questions like credible sets, posterior variance etc involve computing integrals of the form
  \[ \int_{\Theta} f(\theta) \pi(\theta). \]
the Bayesian estimator and computing integrals

- We consider only the most common choice of quadratic loss function
  
  \[ L(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2 \]

- in which case
  
  \[ \hat{\theta}(y_1, \ldots, y_n) = E_{\pi} \theta \]

  so it is the posterior mean.

- So computing the Bayesian estimator is computing the integral wrt the posterior
  
  \[ \int_{\Theta} \theta \pi(\theta) \]

- Similarly answering other inferential questions like credible sets, posterior variance etc involve computing integrals of the form
  
  \[ \int_{\Theta} f(\theta) \pi(\theta). \]
The Monte Carlo Method

\[ I(f) = \int_{\Theta} f(\theta) \pi(\theta) \, d\theta. \]

- Standard Monte Carlo amounts to
  1. sample \( \theta_i \sim \pi \) for \( i = 1, \ldots, k \)
  2. compute \( \hat{I}_k(f) = \frac{1}{k} \sum_i f(\theta_i) \)

- Standard LLN and CLT apply.
- In particular the CLT variance is \( \text{Var}_\pi f \)
- However sampling from \( \pi \) is typically not easy.
The Monte Carlo Method

\[ I(f) = \int_{\Theta} f(\theta) \pi(\theta). \]

- Standard Monte Carlo amounts to
  1. sample \( \theta_i \sim \pi \) for \( i = 1, \ldots, k \)
  2. compute \( \hat{I}_k(f) = \frac{1}{k} \sum_i f(\theta_i) \)
- Standard LLN and CLT apply.
- In particular the CLT variance is \( \text{Var}_\pi f \)
- However sampling from \( \pi \) is typically not easy.
The Monte Carlo Method

\[ I(f) = \int_{\Theta} f(\theta) \pi(\theta). \]

- Standard Monte Carlo amounts to
  1. sample \( \theta_i \sim \pi \) for \( i = 1, \ldots, k \)
  2. compute \( \hat{I}_k(f) = \frac{1}{k} \sum_i f(\theta_i) \)

- Standard LLN and CLT apply.

- In particular the CLT variance is \( Var_{\pi} f \)

- However sampling from \( \pi \) is typically not easy.
The Bayesian setting
Sampling Probability Distributions 1 - direct approaches
Sampling Probability distributions 2 - Markov chains

The Monte Carlo Method

\[ I(f) = \int_{\Theta} f(\theta) \pi(\theta). \]

- Standard Monte Carlo amounts to
  1. sample \( \theta_i \sim \pi \) for \( i = 1, \ldots, k \)
  2. compute \( \hat{I}_k(f) = \frac{1}{k} \sum_i f(\theta_i) \)

- Standard LLN and CLT apply.
- In particular the CLT variance is \( \text{Var}_\pi f \)
- However sampling from \( \pi \) is typically not easy.
The Monte Carlo Method

\[ I(f) = \int_{\Theta} f(\theta) \pi(\theta). \]

- Standard Monte Carlo amounts to
  1. sample \( \theta_i \sim \pi \) for \( i = 1, \ldots, k \)
  2. compute \( \hat{I}_k(f) = \frac{1}{k} \sum_i f(\theta_i) \)

- Standard LLN and CLT apply.
- In particular the CLT variance is \( Var_{\pi} f \)
- However sampling from \( \pi \) is typically not easy.
The Monte Carlo Method

\[ I(f) = \int_{\Theta} f(\theta) \pi(\theta). \]

- Standard Monte Carlo amounts to
  1. sample \( \theta_i \sim \pi \) for \( i = 1, \ldots, k \)
  2. compute \( \hat{I}_k(f) = \frac{1}{k} \sum_i f(\theta_i) \)
- Standard LLN and CLT apply.
- In particular the CLT variance is \( \text{Var}_\pi f \)
- However sampling from \( \pi \) is typically not easy.
Let \( F \) be the cdf of \( \pi \) and define its left continuous inverse version

\[
F^- := \inf\{x : F(x) \geq u\} \quad \text{for} \quad 0 < u < 1.
\]

If \( U \sim U(0, 1) \) then

\[
F^-(U) \sim \pi
\]

Verify the above as an exercise.
Let $F$ be the cdf of $\pi$ and define its left continuous inverse version

$$F^- := \inf\{x : F(x) \geq u\} \quad \text{for} \quad 0 < u < 1.$$ 

If $U \sim U(0, 1)$ then

$$F^-(U) \sim \pi$$

Verify the above as an exercise.
Let $F$ be the cdf of $\pi$ and define its left continuous inverse version

$$F^- := \inf \{ x : F(x) \geq u \} \quad \text{for} \quad 0 < u < 1.$$

If $U \sim U(0, 1)$ then

$$F^-(U) \sim \pi$$

Verify the above as an exercise.
Let $F$ be the cdf of $\pi$ and define its left continuous inverse version

$$F^- := \inf\{x : F(x) \geq u\} \quad \text{for} \quad 0 < u < 1.$$ 

If $U \sim U(0, 1)$ then

$$F^-(U) \sim \pi$$

Verify the above as an exercise.
Let $F$ be the cdf of $\pi$ and define its left continuous inverse version

$$F^{-} := \inf\{x : F(x) \geq u\}$$

for $0 < u < 1$.

If $U \sim U(0, 1)$ then

$$F^{-}(U) \sim \pi$$

Verify the above as an exercise.
Rejection sampling

- Sample $Y$ from density $g(\theta)$ such that

$$\pi(\theta) \leq Cg(\theta) \quad \text{for some} \quad C < \infty$$

- given $Y = \theta$, accept $Y$ with probability

$$\frac{\pi(Y)}{Cg(Y)}$$

- The accepted outcome is distributed as $\pi$

- The average number of trials until acceptance is $1/C$.

- Verify the above as an exercise.
Rejection sampling

- Sample \( Y \) from density \( g(\theta) \) such that

\[
\pi(\theta) \leq C g(\theta)
\]

for some \( C < \infty \).

- given \( Y = \theta \), accept \( Y \) with probability

\[
\frac{\pi(Y)}{C g(Y)}
\]

- The accepted outcome is distributed as \( \pi \).
- The average number of trials until acceptance is \( 1/C \).
- Verify the above as an exercise.
Rejection sampling

- Sample $Y$ from density $g(\theta)$ such that

$$\pi(\theta) \leq Cg(\theta) \quad \text{for some} \quad C < \infty$$

- given $Y = \theta$, accept $Y$ with probability

$$\frac{\pi(Y)}{Cg(Y)}$$

- The accepted outcome is distributed as $\pi$

- The average number of trials until acceptance is $1/C$.

- Verify the above as an exercise.
Rejection sampling

- Sample $Y$ from density $g(\theta)$ such that
  \[ \pi(\theta) \leq C g(\theta) \quad \text{for some} \quad C < \infty \]
- given $Y = \theta$, accept $Y$ with probability
  \[ \frac{\pi(Y)}{C g(Y)} \]
- The accepted outcome is distributed as $\pi$
- The average number of trials until acceptance is $1/C$.
- Verify the above as an exercise.
Rejection sampling

- Sample $Y$ from density $g(\theta)$ such that

\[
\pi(\theta) \leq C g(\theta) \quad \text{for some} \quad C < \infty
\]

- given $Y = \theta$, accept $Y$ with probability

\[
\frac{\pi(Y)}{C g(Y)}
\]

- The accepted outcome is distributed as $\pi$

- The average number of trials until acceptance is $1/C$.

- Verify the above as an exercise.
Importance sampling

▶ Let $g$ be a density such that $\pi(\theta) > 0 \implies g(\theta) > 0$

▶ Then we can write

$$I = \mathbb{E}_\pi f = \int_\Theta f(\theta)\pi(\theta)d\theta = \int_\Theta f(\theta)\frac{\pi(\theta)}{g(\theta)}g(\theta)d\theta$$

$$= \int_\Theta f(\theta)W(\theta)g(\theta)d\theta = \mathbb{E}_g fW.$$

▶ Hence the importance sampling Algorithm:

1. Sample $\theta_i, i = 1, \ldots, k$ iid from $g$
2. Estimate the integral by the unbiased, consistent estimator:

$$\hat{I}_k = \frac{1}{k} \sum_i f(\theta_i)W(\theta_i).$$

▶ Note that compare to iid Monte Carlo the variance of the estimators changes (typically increases) to $\text{Var}(fW)$. 
Importance sampling

- Let $g$ be a density such that $p_i(\theta) > 0 \implies g(\theta) > 0$
- Then we can write

$$I = \mathbb{E}_\pi f = \int_{\Theta} f(\theta)\pi(\theta)d\theta = \int_{\Theta} f(\theta)\frac{\pi(\theta)}{g(\theta)}g(\theta)d\theta$$

$$= \int_{\Theta} f(\theta)W(\theta)g(\theta)d\theta = \mathbb{E}_g fW.$$

- Hence the importance sampling Algorithm:
  1. Sample $\theta_i \ i = 1, \ldots, k$ iid from $g$
  2. Estimate the integral by the unbiased, consistent estimator:

$$\hat{I}_k = \frac{1}{k} \sum_i f(\theta_i)W(\theta_i).$$

- Note that compare to iid Monte Carlo the variance of the estimators changes (typically increases) to $\text{Var}(fW)$. 

Krzysztof Latuszynski (University of Warwick, UK)
Importance sampling

- Let $g$ be a density such that $p_i(\theta) > 0 \implies g(\theta) > 0$
- Then we can write

$$I = \mathbb{E}_\pi f = \int_{\Theta} f(\theta)\pi(\theta) d\theta = \int_{\Theta} f(\theta) \frac{\pi(\theta)}{g(\theta)} g(\theta) d\theta$$

$$= \int_{\Theta} f(\theta) W(\theta) g(\theta) d\theta = \mathbb{E}_g fW.$$

- Hence the importance sampling Algorithm:
  1. Sample $\theta_i, i = 1, \ldots, k$ iid from $g$
  2. Estimate the integral by the unbiased, consistent estimator:

$$\hat{I}_k = \frac{1}{k} \sum_i f(\theta_i) W(\theta_i).$$

- Note that compare to iid Monte Carlo the variance of the estimators changes (typically increases) to $\text{Var}(fW)$. 
Importance sampling

- Let $g$ be a density such that $p_i(\theta) > 0 \implies g(\theta) > 0$
- Then we can write

$$I = \mathbb{E}_\pi f = \int_\Theta f(\theta)\pi(\theta)d\theta = \int_\Theta f(\theta)\frac{\pi(\theta)}{g(\theta)}g(\theta)d\theta$$

$$= \int_\Theta f(\theta)W(\theta)g(\theta)d\theta = \mathbb{E}_g fW.$$

- Hence the importance sampling Algorithm:
  1. Sample $\theta_i$ $i = 1, \ldots, k$ iid from $g$
  2. Estimate the integral by the unbiased, consistent estimator:

$$\hat{I}_k = \frac{1}{k} \sum_i f(\theta_i)W(\theta_i).$$

- Note that compare to iid Monte Carlo the variance of the estimators changes (typically increases) to $\text{Var}(fW)$. 

Krzysztof Latuszynski

University of Warwick, UK
The idea can be extended to a Markov process if the target distribution is of the form

\[ p(\theta_1, \ldots, \theta_n) = p(\theta_1) \prod_{i=2}^{n} p(\theta_i|\theta_{i-1}) \]

We can use a proposal process defined by

\[ q(\theta_1) \quad \text{and} \quad q(\theta_i|\theta_{i-1}). \]
The idea can be extended to a Markov process
if the target distribution is of the form

\[ p(\theta_1, \ldots, \theta_n) = p(\theta_1) \prod_{i=2}^{n} p(\theta_i|\theta_{i-1}) \]

We can use a proposal process defined by

\[ q(\theta_1) \text{ and } q(\theta_i|\theta_{i-1}). \]
The idea can be extended to a Markov process
if the target distribution is of the form

\[ p(\theta_1, \ldots, \theta_n) = p(\theta_1) \prod_{i=2}^{n} p(\theta_i | \theta_{i-1}) \]

We can use a proposal process defined by

\[ q(\theta_1) \quad \text{and} \quad q(\theta_i | \theta_{i-1}). \]
to implement the SIS algorithm:

1. Sample $\theta_1^{(i)} \ i = 1, \ldots, k$ iid from $q$, assign weight

$$w_1^{(i)} = \frac{p(\theta_1^{(i)})}{q(\theta_1^{(i)})}$$

2. For $t = 2, \ldots, n$ simulate

$$\theta_t^{(i)} | \theta_{t-1}^{(i)} \sim q(\theta_t | \theta_{t-1}^{(i)})$$

and update the weight according to

$$w_t^{(i)} = w_{t-1}^{(i)} \frac{p(\theta_t^{(i)} | \theta_{t-1}^{(i)})}{q(\theta_t^{(i)} | \theta_{t-1}^{(i)})}$$

The weakness of importance sampling and SIS is that it is difficult to choose efficient proposal distributions, especially if $\Theta$ is high dimensional.
Markov chains

Let $P$ be a Markov operator on a general state space $\Theta$

So if

$$\theta_0 \sim \nu$$

then for $t = 1, 2, \ldots$

$$\theta_t \sim P(\theta_{t-1}, \cdot)$$

So the distribution of $\theta_1$ is $\nu P$ i.e.

$$\nu P(A) = \int_{\Theta} P(\theta, A) \nu(\theta) d\theta$$

and similarly the distribution of $\theta_t$ is $\nu P^t$ i.e.

$$\nu P^t(A) = \int_{\Theta} P(\theta, A) \nu P^{t-1}(\theta) d\theta$$
Let \( P \) be a Markov operator on a general state space \( \Theta \)

So if

\[ \theta_0 \sim \nu \]

then for \( t = 1, 2, \ldots \)

\[ \theta_t \sim P(\theta_{t-1}, \cdot) \]

So the distribution of \( \theta_1 \) is \( \nu P \) i.e.

\[ \nu P(A) = \int_\Theta P(\theta, A) \nu(\theta) d\theta \]

and similarly the distribution of \( \theta_t \) is \( \nu P^t \) i.e.

\[ \nu P^t(A) = \int_\Theta P(\theta, A) \nu P^{t-1}(\theta) d\theta \]
Markov chains

Let $P$ be a Markov operator on a general state space $\Theta$

So if

$$\theta_0 \sim \nu$$

then for $t = 1, 2, \ldots$

$$\theta_t \sim P(\theta_{t-1}, \cdot)$$

So the distribution of $\theta_1$ is $\nu P$ i.e.

$$\nu P(A) = \int_\Theta P(\theta, A) \nu(\theta) d\theta$$

and similarly the distribution of $\theta_t$ is $\nu P^t$ i.e.

$$\nu P^t(A) = \int_\Theta P(\theta, A) \nu P^{t-1}(\theta) d\theta$$
Under weak assumptions $\nu P^t$ converges as $t \to \infty$ to the same measure, say $\pi_{inv}$ for every $\nu$.

This $\pi_{inv}$ is called stationary or invariant measure and satisfies for every $t$

$$\pi_{inv} P^t = \pi_{inv}$$

So if $t$ is large enough

$$\mathcal{L}(\theta_t) \approx \pi_{inv}$$

STRATEGY: Take the posterior distribution $\pi$ and try to design $P$ so that

$$\pi P = \pi.$$ 

This is feasible more often than you would expect!!!

Under very mild conditions this implies

$$\nu P^t \to \pi \quad \text{for every} \quad \nu.$$ 

We then have for $t$ large enough approximately

$$\theta_t \sim \pi.$$
The Bayesian setting

Sampling Probability Distributions 1 - direct approaches
Sampling Probability distributions 2 - Markov chains

Markov chains

- Under weak assumptions $\nu P^t$ converges as $t \to \infty$ to the same measure, say $\pi_{inv}$ for every $\nu$.
- This $\pi_{inv}$ is called stationary or invariant measure and satisfies for every $t$

$$\pi_{inv} P^t = \pi_{inv}$$

- So if $t$ is large enough

$$\mathcal{L}(\theta_t) \approx \pi_{inv}$$

- STRATEGY: Take the posterior distribution $\pi$ and try to design $P$ so that

$$\pi P = \pi.$$  

- This is feasible more often than you would expect!!!
- Under very mild conditions this implies

$$\nu P^t \to \pi \quad \text{for every} \quad \nu.$$  

- We then have for $t$ large enough approximately

$$\theta_t \sim \pi.$$
Markov chains

- Under weak assumptions $\nu P^t$ converges as $t \to \infty$ to the same measure, say $\pi_{inv}$ for every $\nu$.
- This $\pi_{inv}$ is called stationary or invariant measure and satisfies for every $t$
  \[ \pi_{inv} P^t = \pi_{inv} \]
- So if $t$ is large enough
  \[ \mathcal{L}(\theta_t) \approx \pi_{inv} \]
- STRATEGY: Take the posterior distribution $\pi$ and try to design $P$ so that
  \[ \pi P = \pi. \]
- This is feasible more often than you would expect!!!
- Under very mild conditions this implies
  \[ \nu P^t \to \pi \quad \text{for every} \quad \nu. \]
- We then have for $t$ large enough approximately
  \[ \theta_t \sim \pi. \]
Under weak assumptions $\nu P^t$ converges as $t \to \infty$ to the same measure, say $\pi_{inv}$ for every $\nu$.

This $\pi_{inv}$ is called stationary or invariant measure and satisfies for every $t$

$$\pi_{inv}P^t = \pi_{inv}$$

So if $t$ is large enough

$$\mathcal{L}(\theta_t) \approx \pi_{inv}$$

STRATEGY: Take the posterior distribution $\pi$ and try to design $P$ so that

$$\pi P = \pi.$$ 

This is feasible more often than you would expect!!!

Under very mild conditions this implies

$$\nu P^t \to \pi \quad \text{for every} \quad \nu.$$ 

We then have for $t$ large enough approximately

$$\theta_t \sim \pi.$$
Markov chains

- Under weak assumptions $\nu P^t$ converges as $t \to \infty$ to the same measure, say $\pi_{inv}$ for every $\nu$.
- This $\pi_{inv}$ is called stationary or invariant measure and satisfies for every $t$
  \[ \pi_{inv} P^t = \pi_{inv} \]
- So if $t$ is large enough
  \[ \mathcal{L}(\theta_t) \approx \pi_{inv} \]
- STRATEGY: Take the posterior distribution $\pi$ and try to design $P$ so that
  \[ \pi P = \pi. \]
- This is feasible more often than you would expect!!!
- Under very mild conditions this implies
  \[ \nu P^t \to \pi \quad \text{for every} \quad \nu. \]
- We then have for $t$ large enough approximately
  \[ \theta_t \sim \pi. \]
Markov chains

- Under weak assumptions $\nu P^t$ converges as $t \to \infty$ to the same measure, say $\pi_{inv}$ for every $\nu$.
- This $\pi_{inv}$ is called stationary or invariant measure and satisfies for every $t$
  $$\pi_{inv} P^t = \pi_{inv}$$
- So if $t$ is large enough
  $$\mathcal{L}(\theta_t) \approx \pi_{inv}$$
- STRATEGY: Take the posterior distribution $\pi$ and try to design $P$ so that
  $$\pi P = \pi.$$ 
- This is feasible more often than you would expect!!!
- Under very mild conditions this implies
  $$\nu P^t \to \pi \quad \text{for every} \quad \nu.$$ 
- We then have for $t$ large enough approximately
  $$\theta_t \sim \pi.$$
The approach can be validated asymptotically for estimating

\[ I(f) = \int_\Theta f(\theta) \pi(\theta) d\theta \]

if \( \theta_0, \theta_1, \ldots \) is a Markov chain with dynamics \( P \), then

under very mild conditions LLN holds

\[ \frac{1}{t} \sum_{i=0}^{t-1} f(\theta_i) \to I(f) \]

And also under suitable conditions a CLT holds

\[ \frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} f(\theta_i) \to N(I(f), \sigma_{as}(P,f)) \]

where \( \sigma_{as}(P,f) \) is called asymptotic variance.

There is substantial effort devoted to reliable estimation of \( \sigma_{as}(P,f) \).
The Bayesian setting
Sampling Probability Distributions 1 - direct approaches
Sampling Probability distributions 2 - Markov chains

**CLT for MCMC**

- The approach can be validated asymptotically for estimating

\[
I(f) = \int_{\Theta} f(\theta) \pi(\theta) d\theta
\]

- if \( \theta_0, \theta_1, \ldots \) is a Markov chain with dynamics \( P \), then
  - under very mild conditions LLN holds
    
    \[
    \frac{1}{t} \sum_{i=0}^{t-1} f(\theta_i) \rightarrow I(f)
    \]
  - And also under suitable conditions a CLT holds
    
    \[
    \frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} f(\theta_i) \rightarrow N(I(f), \sigma_{as}(P,f))
    \]

where \( \sigma_{as}(P,f) \) is called asymptotic variance.

- There is substantial effort devoted to reliable estimation of \( \sigma_{as}(P,f) \).
The approach can be validated asymptotically for estimating

\[ I(f) = \int_{\Theta} f(\theta) \pi(\theta) d\theta \]

- if \( \theta_0, \theta_1, \ldots \) is a Markov chain with dynamics \( P \), then
- under very mild conditions LLN holds

\[ \frac{1}{t} \sum_{i=0}^{t-1} f(\theta_i) \to I(f) \]

- And also under suitable conditions a CLT holds

\[ \frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} f(\theta_i) \to N(I(f), \sigma_{as}(P,f)) \]

where \( \sigma_{as}(P,f) \) is called asymptotic variance.

- There is substantial effort devoted to reliable estimation of \( \sigma_{as}(P,f) \).
CLT for MCMC

- The approach can be validated asymptotically for estimating

\[ I(f) = \int_{\Theta} f(\theta) \pi(\theta) d\theta \]

- if \( \theta_0, \theta_1, \ldots \) is a Markov chain with dynamics \( P \), then

- under very mild conditions LLN holds

\[ \frac{1}{t} \sum_{i=0}^{t-1} f(\theta_i) \rightarrow I(f) \]

- And also under suitable conditions a CLT holds

\[ \frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} f(\theta_i) \rightarrow N(I(f), \sigma_{as}(P,f)) \]

where \( \sigma_{as}(P,f) \) is called asymptotic variance.

- There is substantial effort devoted to reliable estimation of \( \sigma_{as}(P,f) \).
The approach can be validated asymptotically for estimating

\[ I(f) = \int_{\Theta} f(\theta) \pi(\theta) d\theta \]

if \( \theta_0, \theta_1, \ldots \) is a Markov chain with dynamics \( P \), then

under very mild conditions LLN holds

\[ \frac{1}{t} \sum_{i=0}^{t-1} f(\theta_i) \rightarrow I(f) \]

And also under suitable conditions a CLT holds

\[ \frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} f(\theta_i) \rightarrow N(I(f), \sigma_{as}(P,f)) \]

where \( \sigma_{as}(P,f) \) is called asymptotic variance.

There is substantial effort devoted to reliable estimation of \( \sigma_{as}(P,f) \).
One way of ensuring $\pi P = \pi$ is the detailed balance condition

$$\pi(\theta_1)P(\theta_1, \theta_2) = \pi(\theta_2)P(\theta_2, \theta_1)$$

In particular consider moving according to some Markov kernel $Q$

i.e. from $\theta_t$ we propose to move to $\theta_{t+1} \sim Q(\theta_t, \cdot)$

And this move is accepted with probability $\alpha(\theta_t, \theta_{t-1})$

Where $\alpha(\theta_t, \theta_{t-1})$ is chosen in such a way that detailed balance holds.

Many such choices for $\alpha(\theta_t, \theta_{t-1})$ are possible

One particular (and optimal in a sense beyond the scope of today) is

$$\alpha(\theta_t, \theta_{t-1}) = \min\{1, \frac{\pi(\theta_{t+1})q(\theta_{t+1}, \theta_t)}{\pi(\theta_t)q(\theta_t, \theta_{t+1})}\}.$$
One way of ensuring $\pi P = \pi$ is the detailed balance condition

$$\pi(\theta_1)P(\theta_1, \theta_2) = \pi(\theta_2)P(\theta_2, \theta_1)$$

In particular consider moving according to some Markov kernel $Q$

- i.e. from $\theta_t$ we propose to move to $\theta_{t+1} \sim Q(\theta_t, \cdot)$
- And this move is accepted with probability $\alpha(\theta_t, \theta_{t-1})$
- Where $\alpha(\theta_t, \theta_{t-1})$ is chosen in such a way that detailed balance holds.
- Many such choices for $\alpha(\theta_t, \theta_{t-1})$ are possible
- One particular (and optimal in a sense beyond the scope of today) is

$$\alpha(\theta_t, \theta_{t-1}) = \min\{1, \frac{\pi(\theta_{t+1})q(\theta_{t+1}, \theta_t)}{\pi(\theta_t)q(\theta_t, \theta_{t+1})}\}.$$
One way of ensuring $\pi P = \pi$ is the detailed balance condition:

$$\pi(\theta_1)P(\theta_1, \theta_2) = \pi(\theta_2)P(\theta_2, \theta_1)$$

In particular consider moving according to some Markov kernel $Q$.

i.e. from $\theta_t$ we propose to move to $\theta_{t+1} \sim Q(\theta_t, \cdot)$

And this move is accepted with probability $\alpha(\theta_t, \theta_{t-1})$

Where $\alpha(\theta_t, \theta_{t-1})$ is chosen in such a way that detailed balance holds.

Many such choices for $\alpha(\theta_t, \theta_{t-1})$ are possible.

One particular (and optimal in a sense beyond the scope of today) is

$$\alpha(\theta_t, \theta_{t-1}) = \min\{1, \frac{\pi(\theta_{t+1})q(\theta_{t+1}, \theta_t)}{\pi(\theta_t)q(\theta_t, \theta_{t+1})}\}.$$
detailed balance and Metropolis Hastings

- One way of ensuring $\pi P = \pi$ is the detailed balance condition

$$
\pi(\theta_1)P(\theta_1, \theta_2) = \pi(\theta_2)P(\theta_2, \theta_1)
$$

- In particular consider moving according to some Markov kernel $Q$
- i.e. from $\theta_t$ we propose to move to $\theta_{t+1} \sim Q(\theta_t, \cdot)$
- And this move is accepted with probability $\alpha(\theta_t, \theta_{t-1})$
- Where $\alpha(\theta_t, \theta_{t-1})$ is chosen in such a way that detailed balance holds.
- Many such choices for $\alpha(\theta_t, \theta_{t-1})$ are possible
- One particular (and optimal in a sense beyond the scope of today) is

$$
\alpha(\theta_t, \theta_{t-1}) = \min\{1, \frac{\pi(\theta_{t+1})q(\theta_{t+1}, \theta_t)}{\pi(\theta_t)q(\theta_t, \theta_{t+1})}\}.
$$
detailed balance and Metropolis Hastings

- One way of ensuring \( \pi P = \pi \) is the detailed balance condition

\[
\pi(\theta_1)P(\theta_1, \theta_2) = \pi(\theta_2)P(\theta_2, \theta_1)
\]

- In particular consider moving according to some Markov kernel \( Q \)
- i.e. from \( \theta_t \) we propose to move to \( \theta_{t+1} \sim Q(\theta_t, \cdot) \)
- And this move is accepted with probability \( \alpha(\theta_t, \theta_{t-1}) \)
- Where \( \alpha(\theta_t, \theta_{t-1}) \) is chosen in such a way that detailed balance holds.
- Many such choices for \( \alpha(\theta_t, \theta_{t-1}) \) are possible
- One particular (and optimal in a sense beyond the scope of today) is

\[
\alpha(\theta_t, \theta_{t-1}) = \min\{1, \frac{\pi(\theta_{t+1})q(\theta_{t+1}, \theta_t)}{\pi(\theta_t)q(\theta_t, \theta_{t+1})}\}.
\]
detailed balance and Metropolis Hastings

- One way of ensuring $\pi P = \pi$ is the detailed balance condition

$$\pi(\theta_1)P(\theta_1, \theta_2) = \pi(\theta_2)P(\theta_2, \theta_1)$$

- In particular consider moving according to some Markov kernel $Q$
- i.e. from $\theta_t$ we propose to move to $\theta_{t+1} \sim Q(\theta_t, \cdot)$
- And this move is accepted with probability $\alpha(\theta_t, \theta_{t-1})$
- Where $\alpha(\theta_t, \theta_{t-1})$ is chosen in such a way that detailed balance holds.
- Many such choices for $\alpha(\theta_t, \theta_{t-1})$ are possible
- One particular (and optimal in a sense beyond the scope of today) is

$$\alpha(\theta_t, \theta_{t-1}) = \min\{1, \frac{\pi(\theta_{t+1})q(\theta_{t+1}, \theta_t)}{\pi(\theta_t)q(\theta_t, \theta_{t+1})}\}.$$
detailed balance and Metropolis Hastings

- One way of ensuring $\pi P = \pi$ is the detailed balance condition

$$\pi(\theta_1) P(\theta_1, \theta_2) = \pi(\theta_2) P(\theta_2, \theta_1)$$

- In particular consider moving according to some Markov kernel $Q$
- i.e. from $\theta_t$ we propose to move to $\theta_{t+1} \sim Q(\theta_t, \cdot)$
- And this move is accepted with probability $\alpha(\theta_t, \theta_{t-1})$
- Where $\alpha(\theta_t, \theta_{t-1})$ is chosen in such a way that detailed balance holds.
- Many such choices for $\alpha(\theta_t, \theta_{t-1})$ are possible
- One particular (and optimal in a sense beyond the scope of today) is

$$\alpha(\theta_t, \theta_{t-1}) = \min\{1, \frac{\pi(\theta_{t+1}) q(\theta_{t+1}, \theta_t)}{\pi(\theta_t) q(\theta_t, \theta_{t+1})}\}.$$
Metropolis-Hastings algorithm

1. Given the current state $\theta_t$ sample the next step proposal

$$\theta_{t+1}^* \sim Q(\theta_t, \cdot)$$

2. Set

$$\theta_{t+1} = \theta_{t+1}^* \quad \text{with probability} \quad \alpha(\theta_t, \theta_{t+1}^*)$$

3. Otherwise set $\theta_{t+1} = \theta_t$.

Exercise: verify the detailed balance for the Metropolis-Hastings algorithm.
The Gibbs Sampler

- For $\Theta = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_d$
- denote the marginals of $\pi$ as $\pi(\theta_k | \theta_{-k})$
- The Gibbs sampler algorithms iterates between updates of $\theta_i | \theta_{-i} \sim \pi(\theta_i | \theta_{-i})$

- Literature: Glynn Asmussen *Stochastic Simulation*
The Gibbs Sampler

- For $\Theta = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_d$

- denote the marginals of $\pi$ as
  \[ \pi(\theta_k | \theta_{-k}) \]

- The Gibbs sampler algorithms iterates between updates of
  \[ \theta_i | \theta_{-i} \sim \pi(\theta_i | \theta_{-i}) \]

- Literature: Glynn Asmussen *Stochastic Simulation*
The Gibbs Sampler

- For $\Theta = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_d$
- denote the marginals of $\pi$ as $\pi(\theta_k | \theta_{-k})$
- The Gibbs sampler algorithms iterates between updates of $\theta_i | \theta_{-i} \sim \pi(\theta_i | \theta_{-i})$
- Literature: Glynn Asmussen *Stochastic Simulation*