Bayesian Statistics
Computational Techniques

Shaobo Jin

Department of Mathematics

Laplace Approximation to Integral

Suppose that we want to approximate the integral

$$\int h(\theta) \exp\{-\ell(\theta)\} d\theta,$$

where θ has the dimension $d \times 1$, p is a known constant, and $\ell(\theta)$ and $h(\theta)$ are smooth functions. If $\ell(\theta)$ is uniquely minimized at $\hat{\theta}$ such that

$$\frac{\partial \ell\left(\hat{\theta}\right)}{\partial \theta} = 0, \qquad \frac{\partial^2 \ell\left(\hat{\theta}\right)}{\partial \theta \partial \theta^T} > 0.$$

Then, the Laplace approximation to the above integral is

$$(2\pi)^{d/2} \sqrt{\det\left(\left[\frac{\partial^2 \ell\left(\hat{\theta}\right)}{\partial \theta \partial \theta^T}\right]^{-1}\right)} \exp\left\{-\ell\left(\hat{\theta}\right)\right\} h\left(\hat{\theta}\right).$$

Laplace Approximation: Example

Example

Suppose that posterior is

$$\begin{aligned} \beta \mid y, \sigma^2 &\sim N\left(\frac{\sum_{i=1}^n y_i}{n+1}, \frac{\sigma^2}{n+1}\right), \\ \sigma^2 \mid y &\sim \text{InvGamma}\left(2+\frac{n}{2}, 2+\frac{1}{2}\left[\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n+1}\right]\right), \end{aligned}$$

where $n = 20$, $\sum_{i=1}^n y_i = 40.4$, and $\sum_{i=1}^n y_i^2 = 93.2$. Approximate $\mathrm{E}[\beta \mid y]$ by Laplace approximation.

Error Analysis

Suppose that we can express $\ell(\theta)$ as $pq(\theta)$ such that the integral is

$$\int h(\theta) \exp\{-pq(\theta)\} d\theta,$$

where $h(\theta)$ and $q(\theta)$ do not include p, and $h(\theta)$ and $q(\theta)$ are smooth functions.

Let $H \overset{\text{def}}{=} \frac{\partial^2 q(\hat{\theta})}{\partial\theta\partial\theta^T} > 0$. The Laplace approximation satisfies

$$(2\pi)^{d/2} \sqrt{\det\left(H^{-1}\left(\hat{\theta}\right)/p\right)} \exp\left\{-pq\left(\hat{\theta}\right)\right\} \left[h\left(\hat{\theta}\right) + O\left(p^{-1}\right)\right].$$

Ratio of Integrals

In practice, we often want to approximate a ratio of integrals such that

$$\int h(\theta)\,\pi(\theta \mid x)\,d\theta = \frac{\int h(\theta)\,f(x \mid \theta)\,\pi(\theta)\,d\theta}{\int f(x \mid \theta)\,\pi(\theta)\,d\theta}.$$

The naive approach is to approximate both the numerator and denominator separately by Laplace approximation and take the ratio of approximations. This yields

$$\int h(\theta)\,\pi(\theta \mid x)\,d\theta \approx h\left(\hat{\theta}\right),$$

which is not recommended.

Moment Generation Function

Consider the moment generation function

$$\begin{aligned} \mathrm{E}\left[\exp\left\{th\left(\theta\right)\right\} \mid x\right] &= \int \exp\left\{th\left(\theta\right)\right\} \pi\left(\theta|x\right) d\theta \\ &= \frac{\int \exp\left\{th\left(\theta\right)\right\} f\left(x \mid \theta\right) \pi\left(\theta\right) d\theta}{\int f\left(x \mid \theta\right) \pi\left(\theta\right) d\theta}. \end{aligned}$$

We apply the Laplace approximation both the denominator and numerator, and take the ratio of approximations. Using the property of the moment generation function,

$$\mathrm{E}\left[h\left(\theta\right) \mid x\right] = \left.\frac{d \log \mathrm{E}\left[\exp\left\{th\left(\theta\right)\right\} \mid x\right]}{dt}\right|_{t=0}.$$

Fully Exponential Laplace Approximation

Let

$$\ell(\theta, t) = -th(\theta) - \log f(x \mid \theta) - \log \pi(\theta).$$

The fully exponential Laplace approximation is

$$\mathrm{E}[h(\theta) \mid x] \approx h\left(\hat{\theta}\right) - \frac{1}{2}\frac{\partial}{\partial t}\log\left|\frac{\partial^2 \ell\left(\tilde{\theta}(t), t\right)}{\partial\theta\partial\theta^T}\right|\Bigg|_{t=0},$$

where $\tilde{\theta}(t)$ maximizes $\ell(\theta, t)$ for a given t, and $\hat{\theta} = \tilde{\theta}(0)$.

Under the same assumptions as for the Laplace approximation, the error rate is $O\left(p^{-2}\right)$.

Expectation Under Posterior

For given data x, we often need to compute the posterior expected value of a function $h(\theta, x)$,

$$\mu(x) = \int h(\theta, x)\, \pi(\theta \mid x)\, d\theta.$$

However, it is not always the case that we can find the closed form expression of $\mu(x)$. Approximations are more often needed.

Suppose that we want to approximate

$$\mathrm{E}[h(x)] = \int h(x)\, f(x)\, dx,$$

where $f(x)$ is the density of random variable/vector X. A natural approximation is to approximate it by the sample mean

$$\bar{h} = \frac{1}{n} \sum_{i=1}^{n} h(x_i).$$

Approximate Expectation by Sample Mean

Under mild conditions, the sample mean

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

has nice properties.

1. Unbiasedness: $\mathrm{E}\left[\bar{h}\right] = \mathrm{E}\left[h(x)\right]$ for any n.
2. Consistency: $\bar{h} \to \mathrm{E}\left[h(x)\right]$ in probability, as $n \to \infty$.
3. Strong consistency: $\bar{h} \to \mathrm{E}\left[h(x)\right]$ almost surely, as $n \to \infty$.
4. Asymptotic normality: $\sqrt{n}\left(\bar{h} - \mathrm{E}\left[h(x)\right]\right) \to N\left(0, \mathrm{Var}\left[h(x)\right]\right)$ in distribution, as $n \to \infty$.

The classic methods (e.g., independent Monte Carlo and importance sampling) have these properties.

Sample From Posterior

For given data x, suppose that we want to compute the posterior expected value

$$\mu(x) = \int h(\theta, x)\, \pi(\theta \mid x)\, d\theta.$$

If $\pi(\theta \mid x)$ is a well-known distribution such that we can easily sample from it, then we draw R independent samples from $\pi(\theta \mid x)$ and the independent Monte Carlo approximation is

$$\hat{\mu}^{\mathrm{IMC}} = \frac{1}{n} \sum_{i=1}^{n} h(\theta_i, x).$$

Independent Monte Carlo: Example

Example

Suppose that posterior is

$$\begin{aligned} \beta \mid y, \sigma^2 &\sim N\left(\frac{\sum_{i=1}^n y_i}{n+1}, \frac{\sigma^2}{n+1}\right), \\ \sigma^2 \mid y &\sim \text{InvGamma}\left(2+\frac{n}{2}, 2+\frac{1}{2}\left[\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n+1}\right]\right), \end{aligned}$$

where $n = 20$, $\sum_{i=1}^n y_i = 40.4$, and $\sum_{i=1}^n y_i^2 = 93.2$. We want to approximate $\mathrm{E}[\beta \mid y]$ by independent Monte Carlo. In this example, we know the true value

$$\mathrm{E}[\beta \mid y] = \frac{\sum_{i=1}^n y_i}{n+1}.$$

Importance Distribution

It is common that it is not straightforward to sample directly from $\pi(\theta \mid x)$. Suppose that it is easy for us to sample directly from another distribution with density $g(\theta \mid x)$ such that $g(\theta \mid x) > 0$ whenever $h(\theta, x)\,\pi(\theta \mid x) \neq 0$.

- We can rewrite $\mu(x)$ as

$$\mu(x) = \int h(\theta, x)\,\frac{\pi(\theta \mid x)}{g(\theta \mid x)}\,g(\theta \mid x)\,d\theta = \mathrm{E}\left[h(\theta, x)\,\frac{\pi(\theta \mid x)}{g(\theta \mid x)} \mid x\right],$$

  where the expectation is taken with respect to $\theta \mid x \sim g(\theta \mid x)$.
- We call $g(\theta \mid x)$ an importance distribution or instrumental distribution.

Importance Sampling Approximation

The importance sampling approximation is

$$\hat{\mu}^{\text{IS}} = \frac{1}{n}\sum_{i=1}^{n} h\left(\theta_i, x\right)\frac{\pi\left(\theta_i \mid x\right)}{g\left(\theta_i \mid x\right)}.$$

Example

Suppose that posterior is

$$\begin{aligned}
\beta \mid y, \sigma^2 &\sim N\left(\frac{\sum_{i=1}^{n} y_i}{n+1}, \frac{\sigma^2}{n+1}\right), \\
\sigma^2 \mid y &\sim \text{InvGamma}\left(2+\frac{n}{2}, 2+\frac{1}{2}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n+1}\right]\right),
\end{aligned}$$

where $n = 20$, $\sum_{i=1}^{n} y_i = 40.4$, and $\sum_{i=1}^{n} y_i^2 = 93.2$. Approximate $\text{E}\left[\sigma^2 \mid y\right]$ by importance sampling.

Normalizing Constant

Since we often derive the posterior using $\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$ by ignoring the normalizing constant, we cannot always evaluate $\pi(\theta \mid x)$.

- It is easy to evaluate $f(x \mid \theta)\pi(\theta)$, but not $m(x)$.
- We can rewrite μ as

$$\mu(x) = \int h(\theta, x)\pi(\theta \mid x)\, d\theta = \frac{\int h(\theta, x) f(x \mid \theta)\pi(\theta)\, d\theta}{\int f(x \mid \theta)\pi(\theta)\, d\theta}.$$

- We can apply the importance sampling trick to both integrals:

$$\int h(\theta, x) f(x \mid \theta)\pi(\theta)\, d\theta = \mathrm{E}\left[h(\theta, x) \underbrace{\frac{f(x \mid \theta)\pi(\theta)}{g(\theta \mid x)}}_{\text{importance weight } w(\theta,x)}\right]$$

where $g(\theta \mid x) > 0$ whenever $\pi(\theta \mid x) \neq 0$, stronger than IS.

Normalized Importance Sampling

The importance sampling approximations to the numerator and denominator are

$$\frac{1}{n}\sum_{i=1}^{n} w(\theta_i, x)\, h(\theta_i, x) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x \mid \theta_i)\,\pi(\theta_i)}{g(\theta_i \mid x)} h(\theta_i, x),$$

$$\frac{1}{n}\sum_{i=1}^{n} w(\theta_i, x) = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x \mid \theta_i)\,\pi(\theta_i)}{g(\theta_i \mid x)}.$$

The ratio is the normalized importance sampling estimator

$$\hat{\mu}^{\mathrm{NIS}} = \frac{\sum_{i=1}^{n} w(\theta_i, x)\, h(\theta_i, x)}{\sum_{i=1}^{n} w(\theta_i, x)}.$$

Normalized Importance Sampling: Example

We can even ignore the constants in $f(y \mid \theta)\pi(\theta)$ in normalized importance sampling.

Example

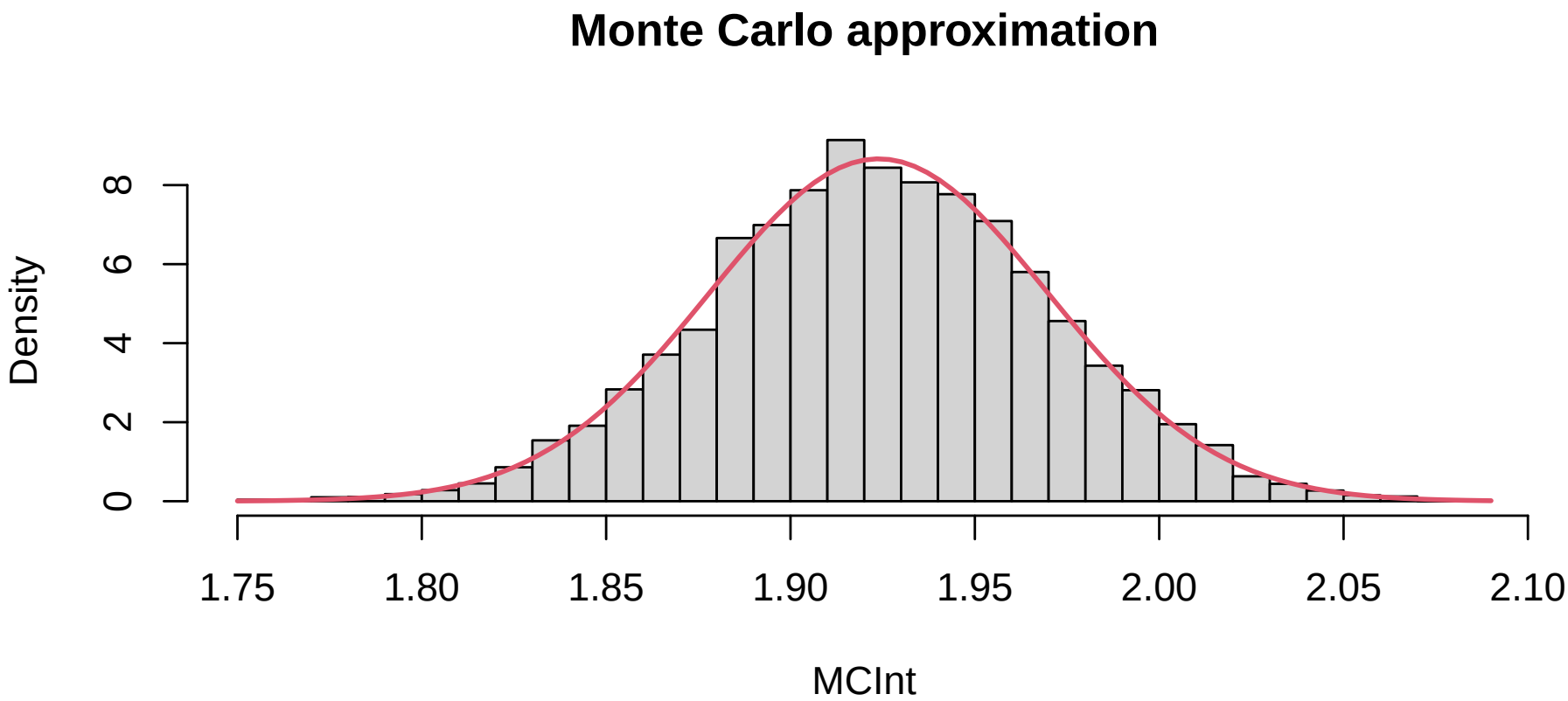
Consider an iid sample of size n from $Y \mid \beta, \sigma^2 \sim N\left(\beta, \sigma^2\right)$. The prior of σ^2 is InvGamma $(2,2)$, and $\beta \mid \sigma^2$ is $N\left(0, \sigma^2\right)$. Then,

$$f(y \mid \theta)\pi(\theta) \quad \propto \quad \frac{\exp\left\{-\frac{(n+1)\beta^2 - 2\beta\sum_{i=1}^n y_i + 4 + \sum_{i=1}^n y_i^2}{2\sigma^2}\right\}}{\left(\sigma^2\right)^{(n+1)/2+3}}$$

We observe $n = 20$, $\sum_{i=1}^n y_i = 40.4$, and $\sum_{i=1}^n y_i^2 = 93.2$. Approximate $\mathrm{E}\left[\sigma^2 \mid y\right]$ by normalized importance sampling.

Randomness

In independent Monte Carlo, importance sampling, and normalized importance sampling, we simulate random numbers from $\pi(\theta \mid x)$ or $g(\theta \mid x)$, then $\hat{\mu}$ is a random variable. This means that we can construct confidence interval for $\hat{\mu}$ using the central limit theorem.



Markov Chain Monte Carlo

We often want to get a sample from the posterior.

- If the posterior follows some well known distribution, we can generate a sample easily.
- If the posterior does not follow any well known distribution, the Markov Chain Monte Carlo (MCMC) is a very popular choice.

The idea of MCMC relies on the Markov property.

Definition

A Markov chain is a sequence of random variables X_i that satisfy the Markov property:

$$\mathrm{P}\left(X_{i+1} \in A \mid X_j = x_j, 0 \leq j \leq i\right) \quad = \quad \mathrm{P}\left(X_{i+1} \in A \mid X_i = x_i\right).$$

Transition Kernel

The transition kernel describes how the Markov chain moves from X_{n-1} to X_n.

- If $\{X_n\}$ is discrete, the transition kernel is a matrix K with elements $\mathrm{P}(X_n = y \mid X_{n-1} = x)$.
- If $\{X_n\}$ is continuous, the Markov property means that

$$\begin{aligned} \mathrm{P}(X_n \in A \mid X_{n-1} = x, \cdots X_0) &= \int_{y \in A} K(x, y)\, dy, \\ f(X_n = y \mid X_{n-1} = x, \cdots X_0) &= f(X_n = y \mid X_{n-1} = x) = K(x, y), \end{aligned}$$

where the transition kernel $K(x, y)$ is the conditional density of Y given $X = x$.

Stationary Distribution

Definition

The distribution p on Ω is a stationary distribution (or invariant distribution) of the Markov chain with the transition kernel K, if

$$\begin{aligned} \mathrm{P}(y) &= \sum_{x \in \mathcal{X}} \mathrm{P}(x) K(x, y), \quad \text{discrete case,} \\ f(y) &= \int_{x \in \mathcal{X}} f(x) K(x, y) \, dx, \quad \text{continuous case,} \end{aligned}$$

where P and f are not generic symbols.

- The stationary distribution means that if the initial state $X_0 \sim \pi(\theta \mid \text{data})$, then $X_n \sim \pi(\theta \mid \text{data})$ for all $n \geq 0$, the same distribution.

Long-Run Property

Theorem

Let $\pi()$ be the stationary distribution of the Markov chain. Under some regularity conditions,

$$\lim_{n\to\infty} \sup_{A} |P(X_n \in A \mid X_0 = x) - \pi(A)| \quad = \quad 0, \ \textit{almost surely},$$

regardless of the initial state $X_0 = x$.

Since the limiting distribution does not depend on the initial state x, the marginal distribution of X_n is approximately the stationary distribution, after large enough iterations.

Choose the Transition Kernel

Our goal is to simulate data from $\pi(\theta \mid x)$. We need to choose the transition kernel K such that the stationary distribution is $\pi(\theta \mid x)$.

Fact

If $\pi(\theta \mid x)$ and $K(\theta, \theta^* \mid x)$ satisfies the detailed balance condition, i.e,

$$K(\theta, \theta^*)\pi(\theta \mid x) \quad = \quad K(\theta^*, \theta)\pi(\theta^* \mid x),$$

for any $\theta, \theta^* \in \Theta$, then $\pi(\theta \mid x)$ is the stationary distribution of the Markov chain with the transition kernel K.

Proposal Distribution

When we simulate random numbers from a Markov chain, we need a proposal distribution

$$T(\theta, \theta^*) = f(\theta^* \mid \theta).$$

Find a proposal distribution $T(\theta, \theta^*)$ that satisfies the detailed balance condition is difficult.

- So with probability $A(\theta, \theta^*)$ we let $\theta^{(n+1)} = \theta^*$ (accept), and probability $1 - A(\theta, \theta^*)$ we let $\theta^{(n+1)} = \theta$ (reject).
- For $\theta^{(n+1)} \neq \theta$, the transition is

$$K(\theta, \theta^*) = T(\theta, \theta^*) A(\theta, \theta^*).$$

Hence, we should seek A such that the detailed balance condition is fulfilled.

Deriving $A(\theta, \theta^*)$

The detailed balance condition is fulfilled, if we choose the acceptance probability to be

$$\begin{aligned} A(\theta, \theta^*) &= \lambda(\theta, \theta^*)\, \pi(\theta^* \mid x)\, T(\theta^*, \theta) \leq 1, \\ A(\theta^*, \theta) &= \lambda(\theta, \theta^*)\, \pi(\theta \mid x)\, T(\theta, \theta^*) \leq 1. \end{aligned}$$

The value λ that maximizes the probability $A(\cdot, \cdot) \leq 1$ is

$$\lambda(\theta, \theta^*) = \min\left\{ \frac{1}{\pi(\theta^* \mid x)\, T(\theta^*, \theta)}, \frac{1}{\pi(\theta \mid x)\, T(\theta, \theta^*)} \right\}.$$

Hence,

$$A(\theta, \theta^*) = \lambda(\theta, \theta^*)\, \pi(\theta^* \mid x)\, T(\theta^*, \theta) = \min\left\{ 1, \frac{\pi(\theta^* \mid x)\, T(\theta^*, \theta)}{\pi(\theta \mid x)\, T(\theta, \theta^*)} \right\}.$$

Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm allows proposal distributions such that $T(\theta, \theta^*) > 0$ if and only if $T(\theta^*, \theta) > 0$.

Algorithm 1: Metropolis-Hastings Algorithm

1 Choose an initial state $\theta^{(0)}$;

2 **for** $t = 1$ *in* $1 : n$ **do**

3 Sample a candidate θ^* from $T\left(\theta^{(t)}, \theta \mid x\right)$;

4 Calculate the ratio $R\left(\theta^{(t)}, \theta^*\right) = \frac{\pi(\theta^*|x)T\left(\theta^*, \theta^{(t)}\right)}{\pi\left(\theta^{(t)}|x\right)T\left(\theta^{(t)}, \theta^*\right)}$;

5 Draw $U \sim \mathrm{U}[0, 1]$;

6 Update

$$\theta^{(t+1)} = \begin{cases} \theta^*, & \text{if } U \leq R\left(\theta^{(t)}, \theta^*\right), \\ \theta^{(t)}, & \text{otherwise.} \end{cases}$$

7 **end**

Metropolis-Hastings Algorithm: Example

Since the ratio $R\left(\theta^{(t)}, \theta^*\right)$ includes $\frac{\pi(\theta^*|x)}{\pi\left(\theta^{(t)}|x\right)}$, we only need to know $\pi\left(\cdot \mid x\right)$ up to a normalizing constant.

Example

Consider an iid sample of size n from $Y \mid \beta, \sigma^2 \sim N\left(\beta, \sigma^2\right)$. The prior of σ^2 is InvGamma $(2, 2)$, and $\beta \mid \sigma^2$ is $N\left(0, \sigma^2\right)$. Then,

$$f\left(y \mid \theta\right)\pi\left(\theta\right) \quad \propto \quad \frac{\exp\left\{-\left[\left(n+1\right)\beta^2 - 2\beta\sum_{i=1}^n y_i + 4 + \sum_{i=1}^n y_i^2\right] / \left(2\sigma^2\right)\right\}}{\left(\sigma^2\right)^{(n+1)/2+3}}.$$

We observe $n = 20$, $\sum_{i=1}^n y_i = 40.4$, and $\sum_{i=1}^n y_i^2 = 93.2$. Obtain a sample from the posterior.

Detailed Balance: Symmetric Proposal

The Metropolis-Hastings algorithm allows asymmetric proposal distributions.

- If the proposal distribution is symmetric, i.e., $T(\theta, \theta^*) = T(\theta^*, \theta)$, then the Metropolis-Hastings algorithm reduces to the Metropolis algorithm.

Example

$\theta^* \mid \theta \sim N\left(\theta, \sigma^2\right)$ is symmetric, since

$$T(\theta, \theta^*) = \frac{1}{\sqrt{2\sigma^2}} \exp\left\{-\frac{(\theta - \theta^*)^2}{2\sigma^2}\right\}.$$

Metropolis Algorithm

Algorithm 2: Metropolis Algorithm

Choose an initial state $\theta^{(0)}$;

for $t = 1$ *in* $1 : n$ **do**

  Sample a candidate θ^* from $T\left(\theta^{(t)}, \theta \mid x\right)$;

  Calculate the ratio $R\left(\theta^{(t)}, \theta^*\right) = \frac{\pi(\theta^* | x)}{\pi\left(\theta^{(t)} | x\right)}$;

  Draw $U \sim \mathrm{U}[0, 1]$;

  Update

$$\theta^{(t+1)} = \begin{cases} \theta^*, & \text{if } U \leq R\left(\theta^{(t)}, \theta^*\right), \\ \theta^{(t)}, & \text{otherwise.} \end{cases}$$

end

Some Examples of Metropolis-Hastings Algorithms

Many different MCMC algorithms differ mainly in how the candidate y is sampled.

- In the random-walk Metropolis algorithm, $\theta^* = \theta^{(t)} + \epsilon$, where ϵ is sampled from some distribution, e.g., Uniform $[-a, a]$, Normal, etc.
- In independence sampler, θ^* is sampled from $g(\cdot)$ that does not depend on $\theta^{(t)}$.
- The Langevin Metropolis-Hastings algorithm explores the shape of the posterior distribution by $\theta^* = \theta^{(t)} + d^{(t)} + \tau\epsilon$, where $\epsilon \sim N(0, I)$ and

$$d^{(t)} = \frac{\tau^2}{2} \frac{\partial \log \pi\left(\theta^{(t)} \mid x\right)}{\partial \theta}.$$

Gibbs Sampler: Conditioning

It can be the case that it is much easier to sample from the conditional distributions than using Metroplis-Hastings from the joint distribution of $\theta \in \Theta \subset \mathbb{R}^d$.

- Suppose that $\theta = (\theta_1, ..., \theta_p)$, where $\theta_i \in \mathbb{R}^{d_i}$.
- Let $\pi_{i|-i}(\theta_i \mid \theta_{-i}, x)$ be the conditional distribution of θ_i given θ_{-i} and x, where $\theta_{-i} = \begin{pmatrix} \theta_1 & \cdots & \theta_{i-1} & \theta_{i+1} & \cdots & \theta_p \end{pmatrix}$.

Algorithm 3: Basic Gibbs Sampler

Choose an initial state $\theta^{(0)}$;
for $t = 1$ *in* $1 : n$ **do**
  for $i = 1$ *in* $1 : p$ **do**
    Draw $\theta_i^{(t+1)} \sim \pi_{i|-i}\left(\theta_i \mid \theta_1^{(t+1)}, ..., \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t)}, ..., \theta_p^{(t)}, x\right)$;
  end
end

Gibbs Sampler: Example

Example

Suppose that our data X_1, ..., X_n are iid from $N\left(\mu, \lambda^{-1}\right)$. The prior distributions of μ and λ are

$$\begin{aligned} \mu &\sim N\left(\mu_0, \lambda_0^{-1}\right), \\ \lambda &\sim \text{Exp}\left(b_0\right). \end{aligned}$$

Use Gibbs sampler to sample random numbers from the posterior distribution of μ, λ.

Why Does Gibbs Sampler Work?

In order to show the Gibbs sampler generate random numbers from the desired stationary distribution, we only need to show

$$\pi\left(\theta^* \mid x\right) \quad = \quad \int K\left(\theta, \theta^*\right) \pi\left(\theta \mid x\right) d\theta,$$

where $\pi\left(\cdot \mid x\right)$ is not a generic symbol.

For simplicity, we consider $p = 2$ and continuous posterior.

- The transition kernel $K\left(\theta, \theta^*\right)$ is

$$K\left(\left(\theta_1, \theta_2\right), \left(\theta_1^*, \theta_2^*\right)\right) \quad = \quad \pi_{1|2}\left(\theta_1^* \mid \theta_2, x\right) \pi_{2|1}\left(\theta_2^* \mid \theta_1^*, x\right).$$

- This transition kernel satisfies

$$\int\int K\left(\left(\theta_1, \theta_2\right), \left(\theta_1^*, \theta_2^*\right)\right) \pi\left(\theta_1, \theta_2 \mid x\right) d\theta_1 d\theta_2 \quad = \quad \pi\left(\theta_1^*, \theta_2^* \mid x\right).$$

Collapsed Gibbs Sampler

Suppose that θ can be partitioned into three groups of parameters $(\theta_1, \theta_2, \theta_3)$.

- The Gibbs sampler samples from the full conditional distributions $\theta_1^{(t+1)} \sim \pi\left(\theta_1 \mid \theta_2^{(t)}, \theta_3^{(t)}, x\right)$, $\theta_2^{(t+1)} \sim \pi\left(\theta_2 \mid \theta_1^{(t+1)}, \theta_3^{(t)}, x\right)$, and $\theta_3^{(t+1)} \sim \pi\left(\theta_3 \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, x\right)$.

In collapsed Gibbs sampler, we can integrate out θ_3 analytically and work with $(\theta_1, \theta_2) \sim \pi(\theta_1, \theta_2 \mid x)$.

- We sample $\theta_1^{(t+1)} \sim \pi\left(\theta_1 \mid \theta_2^{(t)}, x\right)$ and $\theta_2^{(t+1)} \sim \pi\left(\theta_2 \mid \theta_1^{(t+1)}, x\right)$ by Gibbs sampler.
- We then sample $\theta_3^{(t+1)} \sim \pi\left(\theta_3 \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, x\right)$.

Rao-Blackwell Theorem in Statistical Inference

Theorem (Rao-Blackwell Theorem)

Let $\hat{\theta}$ be an unbiased estimator of θ. Suppose that $T = T(X)$ is a sufficient statistic for θ. Then, $\theta^* = E\left[\hat{\theta} \mid T\right]$ is a uniformly minimum variance unbiased estimator of θ, i.e.,

$$Var\left(\hat{\theta}\right) \quad \geq \quad Var\left(\theta^*\right).$$

A weaker version is of the theorem is based on the low of total variance:

$$\mathrm{Var}\left(X\right) \quad = \quad \mathrm{Var}\left(\mathrm{E}\left[X \mid Y\right]\right) + \mathrm{E}\left(\mathrm{Var}\left[X \mid Y\right]\right) \geq \mathrm{Var}\left(\mathrm{E}\left[X \mid Y\right]\right).$$

Rao-Blackwellization

If we are interested in $\mathrm{E}\left[f\left(X,Y\right)\right]$, then

$$\mathrm{Var}\left(f\left(X,Y\right)\right) \quad \geq \quad \mathrm{Var}\left(\mathrm{E}\left[f\left(X,Y\right)\mid Y\right]\right).$$

That is, instead of simulating (X_i, Y_i) to compute $n^{-1}\sum_{i=1}^{n} f\left(X_i, Y_i\right)$, we can simulate only Y_i and compute

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[f\left(X_i,Y_i\right)\mid Y_i\right].$$

This also suggests that we should compute as many analytical steps as possible before Monte Carlo approximation!

Rao-Blackwellization: Example

Example

Consider a Bayesian model, where $X_i \mid \mu, \lambda \sim N\left(\mu, \lambda^{-1}\right)$, $\mu \sim N\left(\mu_0, \lambda_0^{-1}\right)$, and $\lambda \sim \text{Gamma}\left(a_0, b_0\right)$. Then,

$$\mu \mid \lambda, \text{data} \sim N\left(\frac{\lambda_0\mu_0 + \lambda\sum_{i=1}^n x_i}{\lambda_0 + n\lambda}, \frac{1}{\lambda_0 + n\lambda}\right),$$

$$\lambda \mid \mu, \text{data} \sim \text{Gamma}\left(a_0 + \frac{n}{2}, b_0 + \frac{1}{2}\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i\mu + \frac{n}{2}\mu^2\right).$$

We want to approximate $\text{E}\left[\lambda \mid \text{data}\right]$.

Hamiltonian Monte Carlo

- The Metropolis algorithm and the Gibbs sampler often move too slowly through the target distribution when the dimension of the target distribution is high.
- Hamiltonian Monte Carlo (HMC) moves much quicker through the target distribution.
  - For each component in the target distribution, HMC adds a momentum variable and the proposal distribution largely depends on the momentum variable.
  - Both the component in the target distribution and the momentum are updated in the MCMC algorithm.

Hamiltonian Dynamics

The idea of HMC originates from the Hamiltonian dynamics in physics.

- The state of a system consists of the position $\theta \in \mathbb{R}^d$ and the momentum $\phi \in \mathbb{R}^d$ of same dimension.
- The Hamiltonian is a function of θ and ϕ, denoted by $H(\theta, \phi)$.
- The position and the momentum can change over time t. The change is described by the Hamilton's equations:

$$\frac{d\theta_i}{dt} = \frac{\partial H}{\partial \phi_i}, \quad \text{and} \quad \frac{d\phi_i}{dt} = -\frac{\partial H}{\partial \theta_i},$$

for $i = 1, ..., d$.

Potential and Kinetic Energy

For HMC, the Hamiltonian is usually

$$H(\theta, \phi) \quad = \quad U(\theta) + V(\phi),$$

where $U(\theta) = -\log \pi(\theta \mid x)$ is called the potential energy and $V(\phi)$ is called the kinetic energy.

- We want to sample from $\pi(\theta \mid x)$. Hence, ϕ is artificial.
- We often let $\phi \sim N(0, M)$, independent of $\theta \mid x$, for a prespecified covariance matrix M, and $V(\phi)$ the negative log density of ϕ.
- The Hamilton's equations become

$$\frac{d\theta}{dt} = M^{-1}\phi, \quad \text{and} \quad \frac{d\phi}{dt} = \frac{\partial \log \pi(\theta \mid x)}{\partial \theta},$$

arranged as column vectors.

Augmentation

Since θ and ϕ are independent, their joint density is

$$\begin{aligned} f(\theta, \phi \mid x) = \pi(\theta \mid x)\, p(\phi \mid x) &= \exp\{-U(\theta) - V(\phi)\} \\ &= \exp\{-H(\theta, \phi)\}. \end{aligned}$$

We have augmented the problem from sampling θ from $\pi(\theta \mid x)$ to sampling (θ, ϕ) form $\exp\{-H(\theta, \phi)\}$.

1. We first sample ϕ from $N(0, M)$, independent of current θ.
   - Since $\phi \sim N(0, M)$, we already sample ϕ from the desired distribution.
2. We then sample θ, where the new state is proposed by Hamiltonian dynamics by solving the differential equations.

Solve Differential Equation

To solve the differential equations, we consider an approximation known as the leapfrog method. For some stepsize $\epsilon > 0$, we perform half-step updates as

$$\begin{aligned}
\phi\left(t+\frac{\epsilon}{2}\right) &= \phi(t)+\frac{\epsilon}{2}\frac{\partial\phi(t)}{\partial t}=\phi(t)+\frac{\epsilon}{2}\frac{\partial\log\pi(\theta(t)\mid x)}{\partial x},\\
\theta(t+\epsilon) &= \theta(t)+\epsilon\frac{\partial\theta\left(t+\frac{\epsilon}{2}\right)}{\partial t}=\theta(t)+\epsilon M^{-1}\phi\left(t+\frac{\epsilon}{2}\right),\\
\phi(t+\epsilon) &= \phi\left(t+\frac{\epsilon}{2}\right)+\frac{\epsilon}{2}\frac{\partial\phi(t+\epsilon)}{\partial t}=\phi\left(t+\frac{\epsilon}{2}\right)+\frac{\epsilon}{2}\frac{\partial\log\pi(\theta(t+\epsilon)\mid x)}{\partial\theta}.
\end{aligned}$$

Starting from $t=0$, we can get a trajectory at times $\{\epsilon, 2\epsilon, ..., L\epsilon\}$, and approximate the values for $\theta(L\epsilon)$ and $\phi(L\epsilon)$.

Leapfrog Method to Sample θ

Suppose that the current state is (θ, ϕ).

1. Update ϕ with a half-step update by

$$\phi \leftarrow \phi + \frac{\epsilon}{2} \frac{\partial \log \pi(\theta \mid x)}{\partial \theta}.$$

2. For $\ell = 1, ..., L-1$,
   1. Update the position: $\theta \leftarrow \theta + \epsilon M^{-1} \phi$.
   2. Update the momentum:

$$\phi \leftarrow \phi + \epsilon \frac{\partial \log \pi(\theta \mid x)}{\partial \theta}.$$

3. Make one last update on the position: $\theta \leftarrow \theta + \epsilon M^{-1} \phi$.
4. Make one last half-step update of the momentum

$$\phi \leftarrow \phi + \frac{\epsilon}{2} \frac{\partial \log \pi(\theta \mid x)}{\partial \theta}.$$

Metropolis Step

Suppose that the state after such L updates is (θ^*, ϕ^*). We negate the momentum and the new proposal state is $(\theta^*, -\phi^*)$.

- We determine whether to accept the proposal using the Metropolis algorithm, where the acceptance probability is

$$A\left((\theta,\phi),(\theta^*,-\phi^*)\right) = \min\left\{1, \frac{\exp\left\{-H\left(\theta^*,-\phi^*\right)\right\}}{\exp\left\{-H\left(\theta,\phi\right)\right\}}\right\}.$$

- If the proposed state is accepted, then we accept θ^* as a new state for θ, but don't care about ϕ^*.
- No matter we accept or reject the proposal, we will draw a new momentum in the next iteration, independent of previous momentum.

Properties of HMC

Some crucial properties of the Hamiltonian dynamics for MCMC updates include

1. deterministic updates. The Hamiltonian dynamics is deterministic. After running the leapfrog loop L times, we always move the initial state (θ_0, ϕ_0) to the same proposal (θ^*, ϕ^*).
2. reversible. The mapping from the state at time t, denoted by $(\theta(t), \phi(t))$, to the state at time $t+s$, denoted by $(\theta(t+s), \phi(t+s))$, is one-to-one and has an inverse mapping. If we negate the momentum, we will come back from $(\theta(t+s), -\phi(t+s))$ to $(\theta(t), -\phi(t))$.
3. connection between momentum and position. The momentum is changed based on the position since

$$\frac{d\phi_i}{dt} = -\frac{\partial H}{\partial \theta_i} = \frac{\partial \log \pi(\theta \mid x)}{\partial \theta_i}.$$

Tuning Parameters

Tuning of HMC can occur in several places such as

1. the distribution for the momentum,
2. the scaling factor ϵ,
3. the number of leapfrog steps L per iteration.

Some theory suggest that we can tune HMC such that the acceptance probability is around 65%.

No-U-Turn Sampler

The no-U-turn sampler (NUTS) allows us to automatically tune the number of steps L: we increases L until the simulated dynamics is long enough such that the proposed position θ^* starts to move back towards the initial position θ if we run more steps.

- This is measured by the angle between $\theta^* - \theta$ and current momentum ϕ^*.

A basic NUTS works as follows. Given the initial status,

1. Sample $u \mid \theta, \phi \sim \text{Uniform}\left[0, \exp\left\{-H\left(\theta, \phi\right)\right\}\right]$.
2. Apply the leapfrog method (with some modification) until a U-turn occurs.
3. Sample uniformly from the points in $\left\{(\theta, \phi) : \exp\left\{-H\left(\theta, \phi\right)\right\} \geq u\right\}$ that the leapfrog step has visited and the detailed balance condition is fulfilled.

Adaptively Tune ϵ

A too small ϵ will waste computation by taking needlessly tiny steps, and a too large will cause high rejection rates.

- In HMC, we tune ϵ in the warm-up stage of MCMC such that the average acceptance probability δ is the user specified value.
- In NUTS, there is no Metropolis accept/reject step. But we can still compute the ratio as if we were using the accept/reject step and set ϵ such that the pseudo acceptance probability is the user specified value.

In stan, the default is $\delta = 0.8$.

Burn-In Period

The stationary distribution is reached after large enough iterations.

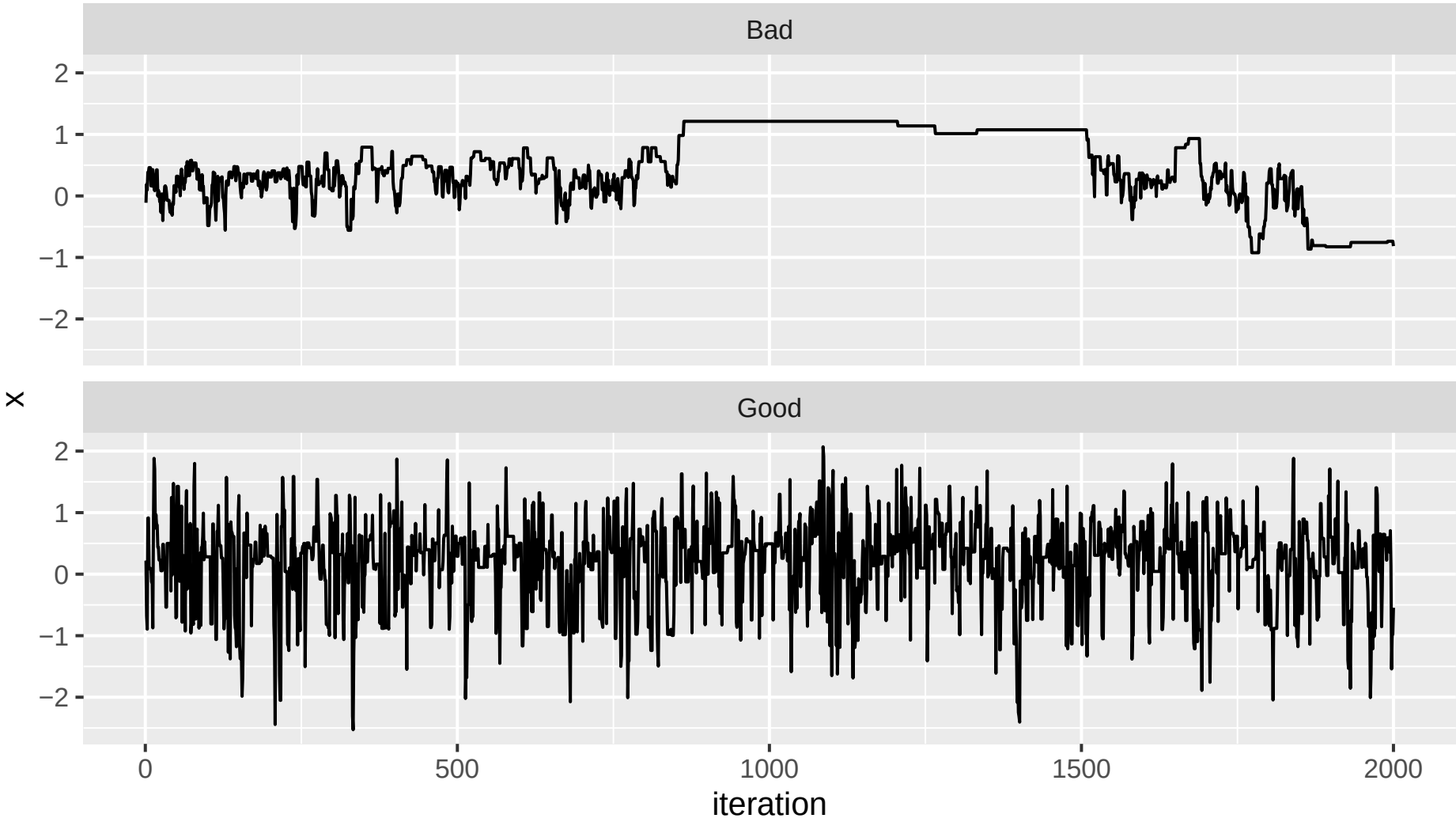
- If the iterations have not proceeded long enough, the simulated numbers may be unrepresentative of the target distribution.

To diminish the influence of the starting values, we can discard the early simulations, known as the burn-in.

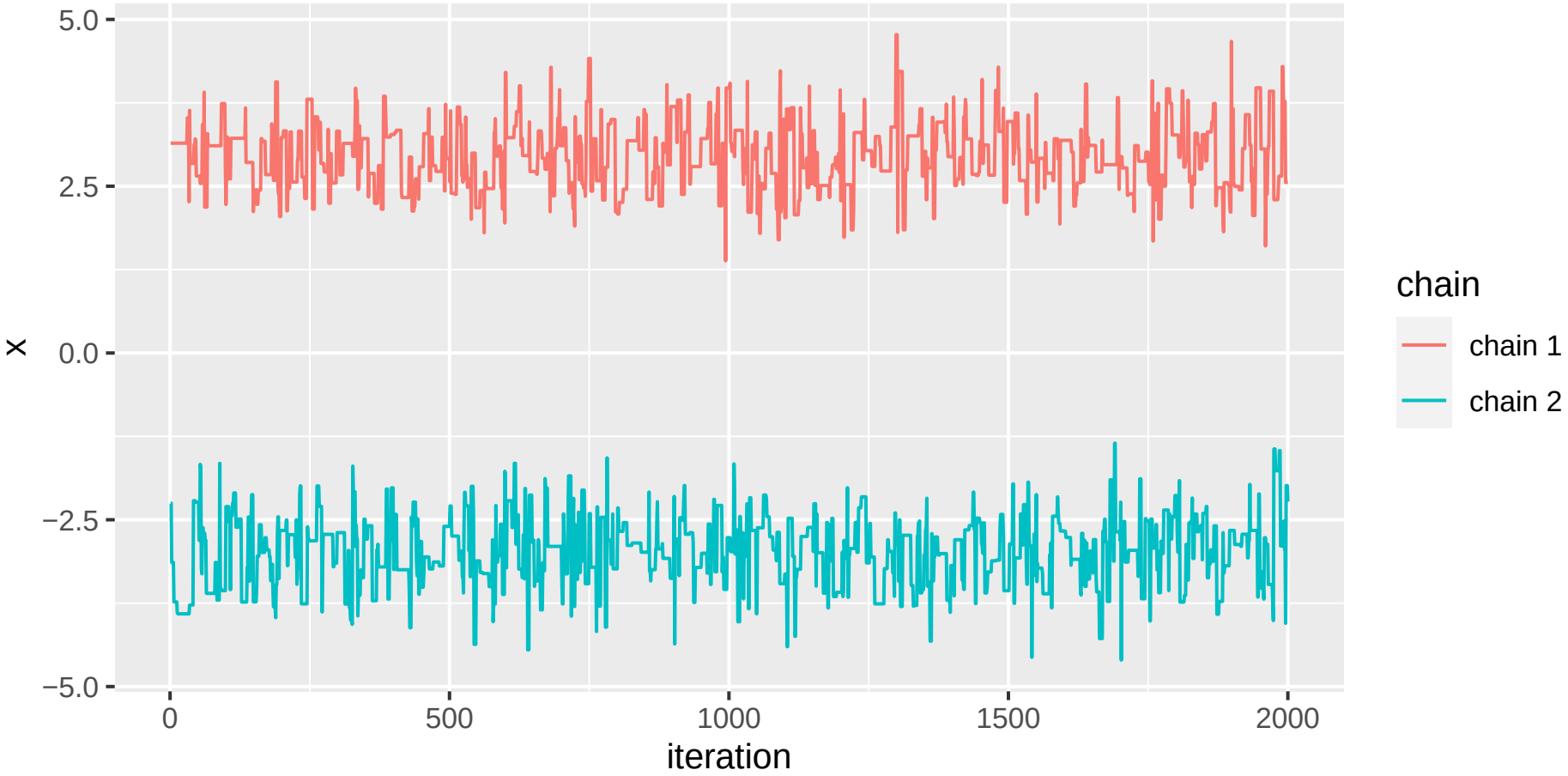
- There is no golden standard on how long the burn-in period should be.
- Hereafter, if the Markov chain has length n, we mean that after the burn-in period, the length is n.

Mixing

We want the Markov chain to show good mixing.



Several Markov Chains



One suggestion is to generate several independent Markov chains, starting from widely separated places.

Gelman-Rubin $\hat{R}$ Statistic: Variation

One way to assess convergence is the Gelman-Rubin $\hat{R}$ statistic. Suppose that we have simulated m chains each with n iterations. Say we have a univariate quantity $y_{ij} = f\left(\theta_j^{(i)}\right)$, where $\theta_j^{(i)}$ is the ith value in the jth chain.

- The variation within the chains is measured by

$$W = \frac{1}{m}\sum_{j=1}^{m}\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_{ij}-\bar{y}_{\cdot j})^2\right],$$

  where $\bar{y}_{\cdot j}$ is the average of $\{y_{ij}\}_{i=1}^{n}$.
- The variation between the chains is measured by

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2,$$

  where $\bar{y}_{\cdot\cdot}$ is the average of all $\bar{y}_{\cdot j}$.

Gelman-Rubin $\hat{R}$ Statistic: Expression

If the Markov chains have reached stationary, then

$$\mathrm{E}[W] = \mathrm{E}[B] = \mathrm{Var}(Y).$$

- We estimate the variance $\mathrm{Var}(Y)$ by

$$\hat{V} = \frac{n-1}{n} W + \frac{1}{n} B.$$

- The Gelman-Rubin $\hat{R}$ statistic is then

$$\hat{R} = \sqrt{\frac{\hat{V}}{W}},$$

  which declines to 1 as $n \to \infty$.
- It is suggested that we keep simulating the Markov chain until $\hat{R} < 1.1$ or even < 1.01.

Variants of Gelman-Rubin $\hat{R}$

Several different versions of $\hat{R}$ have been proposed.

1. One suggestion is to change $\hat{V}$ to

$$\hat{V} = \frac{n-1}{n}W + \frac{1}{n}\left(1+\frac{1}{m}\right)B.$$

   to account for the possibility that $\hat{V}$ is too low.
2. Another suggestion is to split each chain into two parts, yields $2m$ chains of length $n/2$ each. Then compute the $\hat{R}$, pretending that we have simulated $2m$ chains of length $n/2$.
   - This can be useful to detect the case where each chain does not reach stationary but the chains cover a common distribution, e.g, two chains exhibit an X-shape.

Serial Correlation

It is obvious that $\theta^{(t+1)}$ and $\theta^{(t)}$ are not independent draws. Inference from autocorrelated draws is generally less precise than from the same number of independent draws.

- However, such serial correlation is not necessarily a problem. Remember that, at convergence, we reach the stationary distribution.

Algorithm 4: General MCMC Integral

---

1 Sample a Markov chain for a given stationary distribution $\pi(\theta \mid x)$: $\theta^{(1)}$, ..., $\theta^{(R)}$ (after burn-in) ;

2 Approximate $\mu(x)$ by

$$\hat{\mu}^{\text{MCMC}} = \frac{1}{n}\sum_{i=1}^{n} h(\theta_i, x).$$

Long-Run Property

Theorem

Under some conditions, for all starting state $\theta_0 \in \Theta$,

1. *ergodic theorem: For any initial state,*

$$\frac{1}{n}\sum_{i=1}^{n} h(\theta_i, x) \overset{a.s.}{\to} E[h(\theta, x) \mid x] = \mu(x).$$

2. *central limit theorem: Let* $\sigma^2 = Var[h(\theta, x) \mid x]$ *and* $\rho_j = corr\left(h\left(\theta^{(1)}, x\right), h\left(\theta^{(j+1)}, x\right) \mid x\right)$. *Then,*

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} h(\theta_i, x) - \mu(x)\right] \overset{d}{\to} N\left(0, \sigma^2\left(1 + 2\sum_{j=1}^{\infty}\rho_j\right)\right).$$

Effective Sample Size

If we have an iid sample of size n, then

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} h\left(\theta_i, x\right) - \mu\left(x\right)\right] \quad \xrightarrow{d} \quad N\left(0, \sigma^2\right).$$

If we have a converged Markov chain of length n,

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} h\left(\theta_i, x\right) - \mu\left(x\right)\right] \quad \xrightarrow{d} \quad N\left(0,\ \sigma^2\left(1 + 2\sum_{j=1}^{\infty}\rho_j\right)\right).$$

The variance of $\hat{\mu}^{\mathrm{MCMC}}$ is larger than the variance of $\hat{\mu}^{\mathrm{IMC}}$. We define

$$n_{\mathrm{eff}} \quad = \quad \frac{n}{1 + 2\sum_{j=1}^{\infty}\rho_j}$$

as the effective sample size of this Markov chain sample.

Estimate Effective Sample Size

We can also estimate the effective sample size, if we have m Markov chains of length n.

- Following the Gelman-Rubin $\hat{R}$ statistic, we can estimate σ^2 by

$$\hat{V} = \frac{n-1}{n} W + \frac{1}{n} B.$$

- The autocorrelations can be estimated by

$$\hat{\rho}_t = 1 - \frac{\sum_{j=1}^{m} \sum_{i=t+1}^{n} \left(y_{i,j} - y_{i-t,j}\right)^2}{2m\left(n-t\right)\hat{V}}.$$

- The effective sample size is estimated by

$$\hat{n}_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{T} \hat{\rho}_t},$$

where T is the first odd positive integer such that $\hat{\rho}_{T+1} + \hat{\rho}_{T+2}$ is negative.

Alternative Confidence Interval

Suppose that we can divide the Markov chain of length n into b batches (e.g., 20 or proportional to $n^{1/3}$) of m consecutive observations each.

- Let $\bar{y}_j$ be the average of batch j.
- We will treat $\{\bar{y}_j\}$ as iid normal random variables.

An approximate confidence interval is

$$\frac{1}{b}\sum_{j=1}^{b}\bar{y}_j \pm t_{1-\alpha/2}\,(b-1)\sqrt{\frac{1}{b\,(b-1)}\sum_{j=1}^{b}(\bar{y}_j-\bar{y})^2},$$

where $\bar{y}$ is the average of $\{\bar{y}_j\}$. This is just the usual t-confidence interval.

Thinning

Some prefer thinning the sequence by only keeping every kth draw from a sequence in order to reduce serial correlation.

- But whether or not the Markov chain is thinned, it can be used for inferences, provided that it has reached convergence.
- Suppose that the length of the Markov chain is n. We discard $k-1$ out of every k observations and the chain after thinning is n/k.
- Under some assumptions,

$$\begin{aligned} \sqrt{n}\left[\hat{\mu}-\mu\left(x\right)\right] &\xrightarrow{d} N\left(0,\ \tau^2\right), \\ \sqrt{n/k}\left[\hat{\mu}_k-\mu\left(x\right)\right] &\xrightarrow{d} N\left(0,\ \tau_k^2\right), \end{aligned}$$

  where $\hat{\mu}$ and $\hat{\mu}_k$ are the estimators without and with thinning, respectively.
- In fact, it has been proved that, for any $k>1$, $k\tau_k^2>\tau^2$, indicating that discarding $k-1$ out of every k observations will increase the variance.

Simulation Under Posterior

Using MCMC and other methods, we can simulate n random numbers from the posterior distribution $\pi(\theta \mid x)$. Using the simulated θ, we can

1. approximate the posterior mean: $n^{-1}\sum_{i=1}^{n}\theta^{(i)} \to \mathrm{E}[\theta \mid x]$.
2. approximate the posterior probability:

$$\frac{1}{n}\sum_{i=1}^{n} 1\left(\theta^{(i)} \in A\right) \quad \to \quad \mathrm{E}[1(\theta \in A) \mid x] = \mathrm{P}(\theta \in A \mid x).$$

3. approximate predictive density:

$$\frac{1}{n}\sum_{i=1}^{n} f\left(x_{\text{new}} \mid x, \theta^{(i)}\right) \quad \to \quad \int f(x_{\text{new}} \mid x, \theta)\,\pi(\theta \mid x)\,d\theta.$$

4. approximate mean of predictive distribution: $n^{-1}\sum_{i=1}^{n} x_{\text{new}}^{(i)} \to \mathrm{E}[x_{\text{new}} \mid x]$, where $x_{\text{new}}^{(i)}$ is simulated from $f\left(x_{\text{new}} \mid x, \theta^{(i)}\right)$.

Approximate Posterior

If the posterior distribution family is difficult to handle, it can be useful to approximate it by another distribution family that is easier to handle.

- The Kullback-Leibler divergence for distributions P and Q with respective densities p and q are

$$\mathrm{KL}(q,p) = \int q(\theta) \log\left[\frac{q(\theta)}{p(\theta)}\right] d\theta \geq 0.$$

- We choose a model $\mathcal{D}$ for the posterior, called the variational family.
- The variational density is

$$q^*(\theta \mid x) = \arg\min_{q \in \mathcal{D}} \mathrm{KL}(q(\theta \mid x),\ \pi(\theta \mid x)).$$

Variational Bayesian Inference

The idea of variational inference (VI) is to use $q^*(\theta \mid x) \in \mathcal{D}$ instead of $\pi(\theta \mid x)$ and to explore the properties of $\mathcal{D}$.

We need to choose $\mathcal{D}$ ourselves.

- Trade-off: too simple $\mathcal{D}$ poorly approximates $\pi(\theta \mid x)$ but too complex $\mathcal{D}$ is hard to handle.
- One choice is the mean-field variational family $\mathcal{D}_{\text{MF}}$, where

$$q(\theta \mid x) = \prod_{j=1}^{m} q_j(\theta_j \mid x),$$

that is, the components in θ are independent. We call $q_j(\theta_j \mid x)$ the jth variational factor.

Evidence Lower Bound

The Kullback-Leilber divergence satisfies

$$\mathrm{KL}\left(q\left(\theta \mid x\right),\ \pi\left(\theta \mid x\right)\right) \quad = \quad \log\left[m\left(x\right)\right] - \underbrace{\int q\left(\theta \mid x\right) \log\left[\frac{p\left(\theta, x\right)}{q\left(\theta \mid x\right)}\right] d\theta}_{\text{evidence lower bound ELBO}(q)}.$$

Since $\mathrm{KL}\left(q\left(\theta \mid x\right),\ \pi\left(\theta \mid x\right)\right) \geq 0$, the ELBO satisfies

$$\mathrm{ELBO}\left(q\right) \quad \leq \quad \log\left[m\left(x\right)\right],$$

a lower bound of the log-marginal likelihood of x.

- Minimizing the KL divergence is the same as maximization of ELBO.

Variational Inference in Linear Regression

Example

Suppose that $y \mid \beta \sim N_n(X\beta, \Sigma)$ and $\beta \sim N_p\left(\mu_0, \Lambda_0^{-1}\right)$, where Σ is known. The posterior is $\beta \mid y \sim N\left(\mu_n, \Lambda_n^{-1}\right)$, where

$$\begin{aligned} \Lambda_n &= \Lambda_0 + X^T\Sigma^{-1}X, \\ \mu_n &= \Lambda_n^{-1}\left(\Lambda_0\mu_0 + X^T\Sigma^{-1}y\right). \end{aligned}$$

Consider the mean-field variational family

$$\mathcal{D}_{\text{MF}} = \left\{N_p\left(\mu, \Lambda^{-1}\right) : \ \mu \in \mathbb{R}^p, \Lambda \text{ is diagonal}\right\}.$$

Find the ELBO and the variational density.

Explicit Expression of $\mathcal{D}_{\text{MF}}$

Theorem

Consider the mean field variational family $\mathcal{D}_{MF}$, where

$$q(\theta \mid x) = \prod_{j=1}^{m} q_j(\theta_j \mid x).$$

Let θ_k be the kth group in θ and

$$q_k^*(\theta_k \mid x) = \arg\min_{q_k} KL(q(\theta \mid x),\ \pi(\theta \mid x)).$$

Then,

$$q_k(\theta_k \mid x) \propto \exp\left\{\int q_{-k}(\theta_{-k} \mid x) \log \pi(\theta_k \mid \theta_{-k}, x)\, d\theta_{-k}\right\}.$$

Coordinate Ascent Variational Inference Algorithm

The previous theorem suggests the following stepwise conditioning to approximate $q^*(\theta \mid x)$.

Algorithm 5: Coordinate ascent variational inference (CAVI) Algorithm

Choose an initial approximation $\hat{q}^{(0)}(\theta \mid x) = \prod_{j=1}^{m} \hat{q}_j^{(0)}(\theta_j \mid x)$;
for $t = 1$ *in* $1 : T$ **do**
  for $j = 1$ *in* $1 : m$ **do**
    Calculate $\hat{q}_j^{(t)}(\theta_j \mid x) \propto \exp\left\{\int q_{-j}(\theta \mid x) \log \pi(\theta_j \mid \theta_{-j}, x)\, d\theta_{-j}\right\}$, where

$$q_{-j}(\theta \mid x) = \left[\prod_{k=1}^{j-1} \hat{q}_k^{(t)}(\theta_k \mid x)\right]\left[\prod_{k=j+1}^{m} \hat{q}_k^{(t-1)}(\theta_k \mid x)\right]$$

  end
end

CAVI Algorithm: Example

Example

Suppose that we have an iid sample $X_i \mid \mu, \sigma^2 \sim N\left(\mu, \sigma^2\right)$, $i = 1, ..., n$.
The priors are $\mu \mid \sigma^2 \sim N\left(\mu_0, \sigma^2/\lambda_0\right)$ and $\sigma^2 \sim \text{InvGamma}\left(a_0, b_0\right)$.
The posterior is

$$\mu \mid x, \sigma^2 \sim N\left(\mu_n,\ \sigma^2/\lambda_n\right), \quad \text{and} \quad \sigma^2 \mid x \sim \text{InvGamma}\left(a_n, b_n\right).$$

where $\lambda_n = \lambda_0 + n$, $\mu_n = \lambda_n^{-1}\left(\lambda_0\mu_0 + \sum_{i=1}^n x_i\right)$, $a_n = a_0 + \frac{n}{2}$, and

$$b_n \quad = \quad b_0 + \frac{1}{2}\left(\sum_{i=1}^n x_i^2 + \lambda_0\mu_0^2 - \lambda_n\mu_n^2\right).$$

Stan

Stan is a c++ library for Bayesian inference using HMC to obtain posterior simulations.

- Rstan is the R interface to Stan.
- PyStan is the Python interface to Stan.

It is the state-of-the-art library for doing Bayesian statistics.

A Stan model consists of

1. data,
2. parameters,
3. statistical model.

R Package rstanarm

The R package rstanarm emulates the R syntax but uses Stan via the rstan package to fit models in the background. So you skip writing the Stan syntax.

- Various common regression models have been implemented in rstanarm.
- Another benefit is that various visualization tools in R can be used.