Introduction to Bayesian Computation

Dr. Jarad Niemi

STAT 544 - Iowa State University

March 26, 2019

Bayesian computation

Goals:

- $E_{\theta|y}[h(\theta)|y] = \int h(\theta)p(\theta|y)d\theta$
- $p(y) = \int p(y|\theta)p(\theta)d\theta = E_\theta[p(y|\theta)]$

Approaches:

- Deterministic approximation
- Monte Carlo approximation
  - Theoretical justification
  - Gridding
  - Inverse CDF
  - Accept-reject

Numerical integration

- Deterministic methods where

$$E[h(\theta)|y] = \int h(\theta)p(\theta|y)d\theta \approx \sum_{S=1}^{S} w_s h\left(\theta^{(s)}\right) p\left(\theta^{(s)}\middle| y\right)$$

  and
  - $\theta^{(s)}$ are selected points,
  - w_s is the weight given to the point $\theta^{(s)}$, and
  - the error can be bounded.
- Monte Carlo (simulation) methods where

$$E[h(\theta)|y] = \int h(\theta)p(\theta|y)d\theta \approx \sum_{S=1}^{S} w_s h\left(\theta^{(s)}\right)$$

  and
  - $\theta^{(s)} \stackrel{ind}{\sim} g(\theta)$ (for some proposal distribution g),
  - $w_s = p(\theta^{(s)}|y)/g(\theta^{(s)})$,
  - and we have SLLN and CLT.

Example: Normal-Cauchy model

Let $Y \sim N(\theta, 1)$ with $\theta \sim Ca(0, 1)$. The posterior is

$$p(\theta|y) \propto p(y|\theta)p(\theta) \propto \frac{\exp(-(y-\theta)^2/2)}{1+\theta^2} = q(\theta|y)$$

which is not a known distribution. We might be interested in

1. normalizing this posterior, i.e. calculating

$$c(y) = \int q(\theta|y)d\theta$$

2. or in calculating the posterior mean, i.e.

$$E[\theta|y] = \int \theta p(\theta|y)d\theta = \int \theta \frac{q(\theta|y)}{c(y)}d\theta.$$

Normal-Cauchy: marginal likelihood

```r
y = 1 # Data
```

```r
q = function(theta, y, log = FALSE) {
  out = -(y-theta)^2/2-log(1+theta^2)
  if (log) return(out)
  return(exp(out))
}

# Find normalizing constant for q(theta|y)
w = 0.1
theta = seq(-5,5,by=w)+y
(cy = sum(q(theta,y)*w))                  # gridding based approach
```

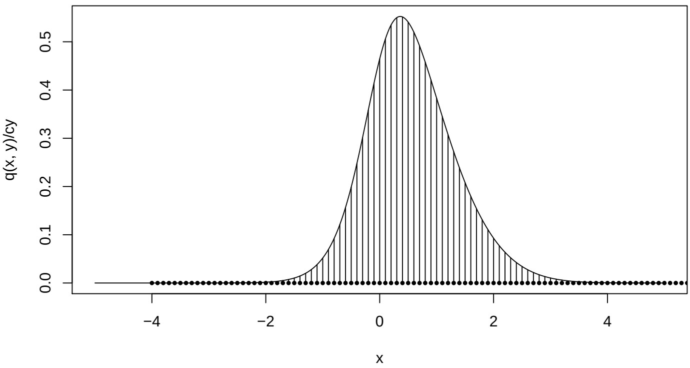
```
[1] 1.305608
```

```r
integrate(function(x) q(x,y), -Inf, Inf) # numerical integration
```

```
1.305609 with absolute error < 0.00013
```

Normal-Cauchy: distribution

```
curve(q(x,y)/cy, -5, 5, n=1001)
points(theta,rep(0,length(theta)), cex=0.5, pch=19)
segments(theta,0,theta,q(theta,y)/cy)
```



Posterior expectation

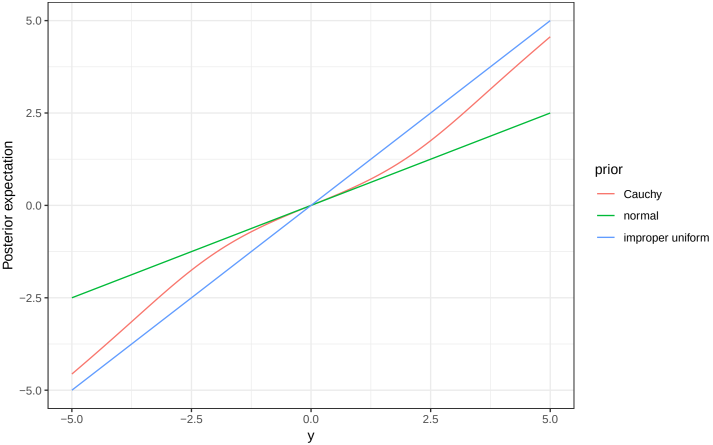
$$E[h(\theta)|y] \approx \sum_{s=1}^{S} w_s h\left(\theta^{(s)}\right) p\left(\theta^{(s)}\middle| y\right) = \sum_{s=1}^{S} w_s h\left(\theta^{(s)}\right) \frac{q\left(\theta^{(s)}\middle| y\right)}{c(y)}$$

```
h = function(theta) theta

sum(w*h(theta)*q(theta,y)/cy)


[1] 0.5542021
```

Posterior expectation as a function of observed data



Convergence review

Three main notions of convergence of a sequence of random variables $X_1, X_2, \ldots$ and a random variable X:

- Convergence in distribution ($X_n \stackrel{d}{\to} X$):
$$\lim_{n\to\infty} F_n(X) = F(x).$$
- Convergence in probability (WLLN, $X_n \stackrel{p}{\to} X$):
$$\lim_{n\to\infty} P(|X_n - X| \ge \epsilon) = 0.$$
- Almost sure convergence (SLLN, $X_n \xrightarrow{a.s.} X$):
$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Implications:

- Almost sure convergence implies convergence in probability.
- Convergence in probability implies convergence in distribution.

Here,

- X_n will be our approximation to an integral and X the true (constant) value of that integral or
- X_n will be a standardized approximation and X will be $N(0,1)$.

Monte Carlo integration

Consider evaluating the integral

$$E[h(\theta)] = \int_\Theta h(\theta)p(\theta)d\theta$$

using the Monte Carlo estimate

$$\hat{h}_S = \frac{1}{S}\sum_{s=1}^{S} h\left(\theta^{(s)}\right)$$

where $\theta^{(s)} \stackrel{ind}{\sim} p(\theta)$. We know

- SLLN: $\hat{h}_S \xrightarrow{a.s.} E[h(\theta)]$.
- CLT: if h^2 has finite expectation, then

$$\frac{\hat{h}_S - E[h(\theta)]}{\sqrt{v_S/S}} \xrightarrow{d} N(0,1)$$

  where

$$v_S = Var[h(\theta)] \approx \frac{1}{S}\sum_{s=1}^{S}\left[h\left(\theta^{(s)}\right) - \hat{h}_S\right]^2$$

  or any other consistent estimator.

Definite integral

Suppose you are interested in evaluating

$$\mathrm{I} = \int_0^1 e^{-\theta^2/2} d\theta.$$

Then set

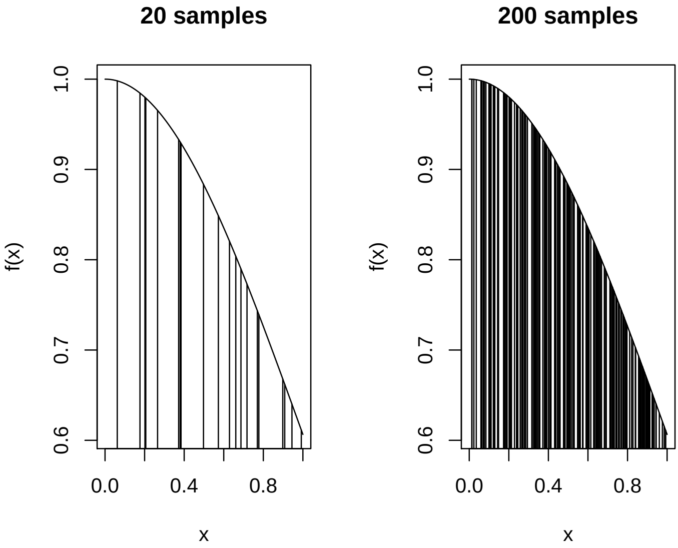
- $h(\theta) = e^{-\theta^2/2}$ and
- $p(\theta) = 1$, i.e. $\theta \sim \text{Unif}(0,1)$.

and approximate by a Monte Carlo estimate via

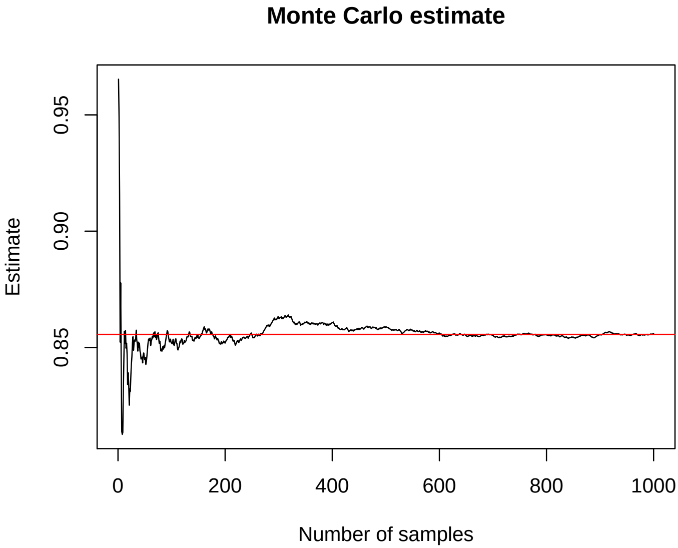
1. For $s = 1, \ldots, S$,
   a. sample $\theta^{(s)} \stackrel{ind}{\sim} Unif(0,1)$ and
   b. calculate $h\left(\theta^{(s)}\right)$.
2. Calculate

$$\mathrm{I} \approx \frac{1}{S} \sum_{s=1}^{S} h(\theta^{(s)}).$$

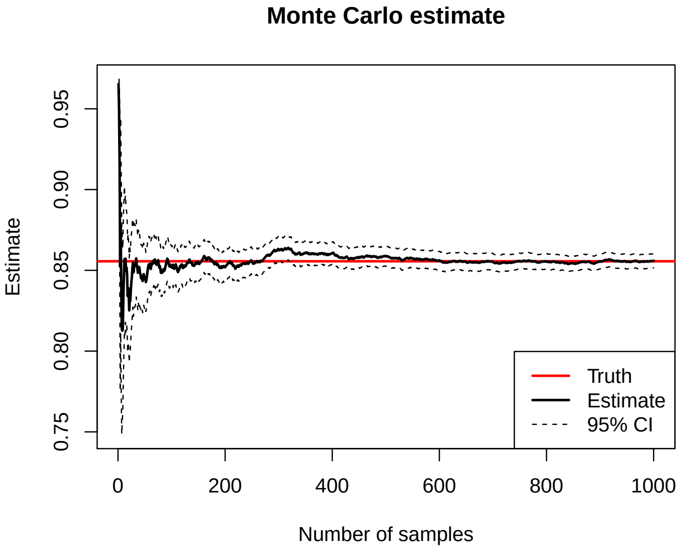
Monte Carlo sampling randomly infills



Strong law of large numbers



Central limit theorem



Infinite bounds

Suppose $\theta \sim N(0,1)$ and you are interested in evaluating

$$E[\theta] = \int_{-\infty}^{\infty} \theta \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2} d\theta$$

Then set

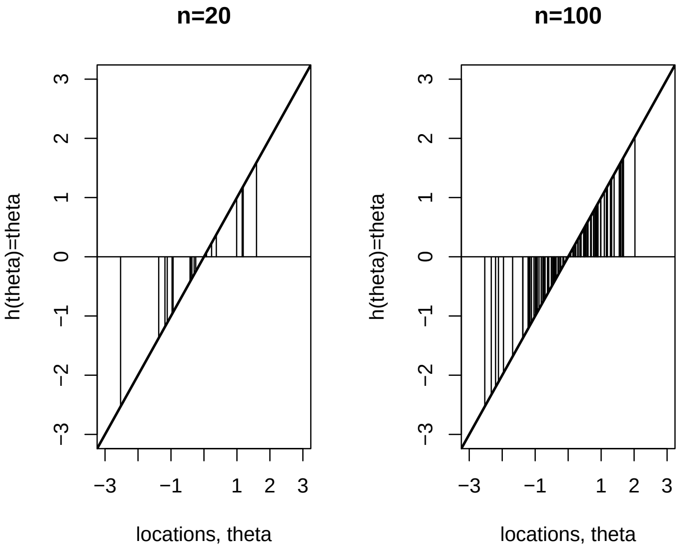
- $h(\theta) = \theta$ and
- $g(\theta) = \phi(\theta)$, i.e. $\theta \sim N(0,1)$.

and approximate by a Monte Carlo estimate via

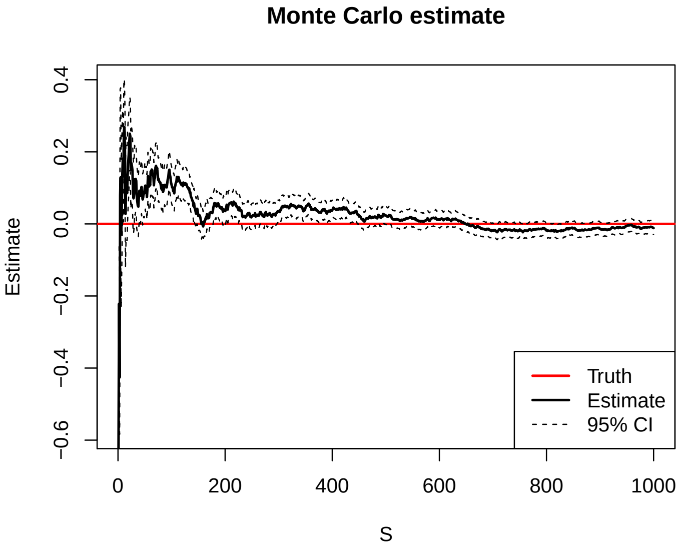
1. For $s = 1, \ldots, S$,
   a. sample $\theta^{(s)} \stackrel{ind}{\sim} N(0,1)$ and
   b. calculate $h\left(\theta^{(s)}\right)$.
2. Calculate

$$E[\theta] \approx \frac{1}{S} \sum_{s=1}^{S} h(\theta^{(s)}).$$

Non-uniform sampling



Monte Carlo estimate



Monte Carlo approximation via gridding

Rather than determining $c(y)$ and then $E[\theta|y]$ via deterministic gridding (all w_i are equal), we can use the grid as a discrete approximation to the posterior, i.e.

$$p(\theta|y) \approx \sum_{i=1}^{N} p_i \delta_{\theta_i}(\theta) \qquad p_i = \frac{q(\theta_i|y)}{\sum_{s=1}^{N} q(\theta_j|y)}$$

where $\delta_{\theta_i}(\theta)$ is the Dirac delta function, i.e.

$$\delta_{\theta_i}(\theta) = 0 \,\forall\, \theta \neq \theta_i \qquad \int \delta_{\theta_i}(\theta) d\theta = 1.$$

This discrete approximation to $p(\theta|y)$ can be used to approximate the expectation $E[h(\theta)|y]$ deterministically or via simulation, i.e.

$$E[h(\theta)|y] \approx \sum_{i=1}^{N} p_i h(\theta_i) \qquad E[h(\theta)|y] \approx \frac{1}{S}\sum_{s=1}^{S} h\left(\theta^{(s)}\right)$$

where $\theta^{(s)} \stackrel{ind}{\sim} \sum_{i=1}^{N} p_i \delta_{\theta_i}(\theta)$ (with replacement).

Example: Normal-Cauchy model

```
y = 1 # Data

# Small number of grid locations
theta = seq(-5,5,length=1e2+1)+y; p = q(theta,y)/sum(q(theta,y)); sum(p*theta)

[1] 0.5542021

mean(sample(theta,prob=p,replace=TRUE))

[1] 0.6118812

# Large number of grid locations
theta = seq(-5,5,length=1e6+1)+y; p = q(theta,y)/sum(q(theta,y)); sum(p*theta)

[1] 0.5542021

mean(sample(theta,1e2,prob=p,replace=TRUE)) # But small MC sample

[1] 0.598394

# Truth
post_expectation(1)

[1] 0.5542021
```

Inverse cumulative distribution function

Definition

The cumulative distribution function (cdf) of a random variable X is defined by

$$F_X(x) = P_X(X \le x) \qquad \text{for all } x.$$

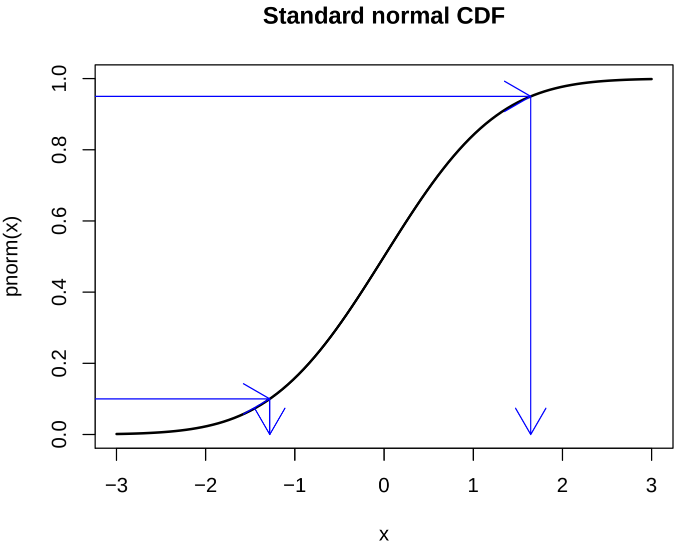
Lemma

Let X be a random variable whose cdf is $F(x)$ and you have access to the inverse cdf of X, i.e. if

$$u = F(x) \quad \implies \quad x = F^{-1}(u).$$

If $U \sim Unif(0,1)$, then $X = F^{-1}(U)$ is a simulation from the distribution for X.

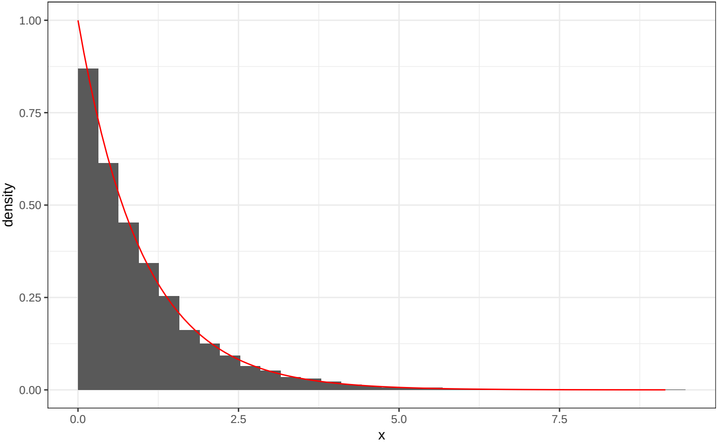
Inverse CDF



Exponential example

For example, to sample $X \sim Exp(1)$,

1. Sample $U \sim Unif(0, 1)$.
2. Set $X = -\log(1 - U)$, or $X = -\log(U)$.



Sampling from a univariate truncated distribution

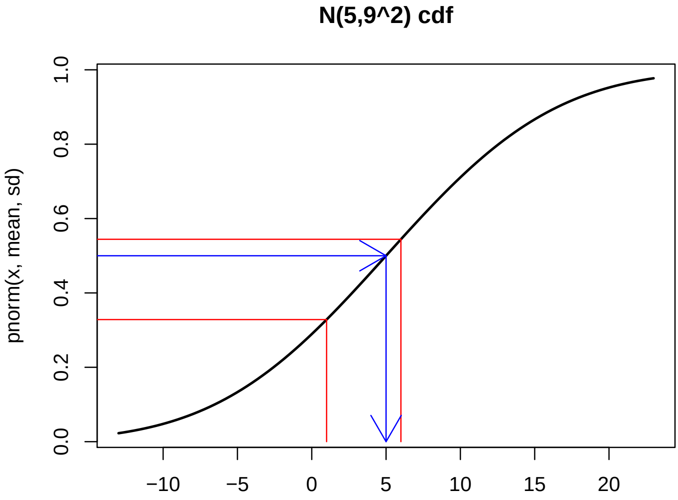
Suppose you wish to sample from $X \sim N(\mu, \sigma^2)\mathrm{I}(a < X < b)$, i.e. a normal random variable with untruncated mean μ and variance σ^2, but truncated to the interval (a, b). Suppose the untruncated cdf is F and inverse cdf is F^{-1}.

1. Calculate endpoints $p_a = F(a)$ and $p_b = F(b)$.
2. Sample $U \sim Unif(p_a, p_b)$.
3. Set $X = F^{-1}(U)$.

This just avoids having to recalculate the normalizing constant for the pdf, i.e. $1/(F^{-1}(b) - F^{-1}(a))$.

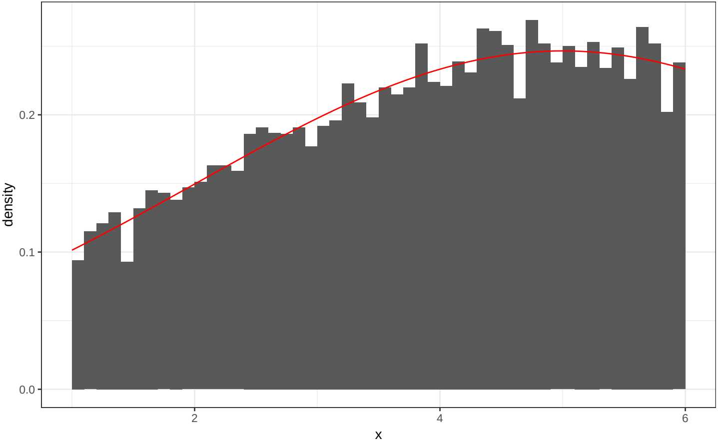
Truncated normal

$X \sim N(5, 9^2)\mathrm{I}(1 \leq X \leq 6)$



Truncated normal

$X \sim N(5, 9^2)\mathrm{I}(1 \le X \le 6)$



Rejection sampling

Suppose you wish to obtain samples $\theta \sim p(\theta|y)$, rejection sampling performs the following

1. Sample a proposal $\theta^* \sim g(\theta)$ and $U \sim Unif(0,1)$.
2. Accept $\theta = \theta^*$ as a draw from $p(\theta|y)$ if $U \le p(\theta^*|y)/Mg(\theta^*)$, otherwise return to step 1.

where M satisfies $M\,g(\theta) \ge p(\theta|y)$ for all θ.

- For a given proposal distribution $g(\theta)$, the optimal M is $M = \sup_\theta p(\theta|y)/g(\theta)$.
- The probability of acceptance is $1/M$.

The accept-reject idea is to create an envelope, $M\,g(\theta)$, above $p(\theta|y)$.

Rejection sampling with unnormalized density

Suppose you wish to obtain samples $\theta \sim p(\theta|y) \propto q(\theta|y)$, rejection sampling performs the following

1. Sample a proposal $\theta^* \sim g(\theta)$ and $U \sim Unif(0,1)$.
2. Accept $\theta = \theta^*$ as a draw from $p(\theta|y)$ if $U \le q(\theta^*|y)/M^\dagger g(\theta^*)$, otherwise return to step 1.

where $M^\dagger$ satisfies $M^\dagger\, g(\theta) \ge q(\theta|y)$ for all θ.

- For a given proposal distribution $g(\theta)$, the optimal $M^\dagger$ is $M^\dagger = \sup_\theta q(\theta|y)/g(\theta)$.
- The acceptance probability is $1/M = c(y)/M^\dagger$.

The accept-reject idea is to create an envelope, $M^\dagger\, g(\theta)$, above $q(\theta|y)$.

Example: Normal-Cauchy model

If $Y \sim N(\theta, 1)$ and $\theta \sim Ca(0, 1)$, then

$$p(\theta|y) \propto e^{-(y-\theta)^2/2} \frac{1}{(1+\theta^2)}$$

for $\theta \in \mathbb{R}$.

Choose a $N(y, 1)$ as a proposal distribution, i.e.

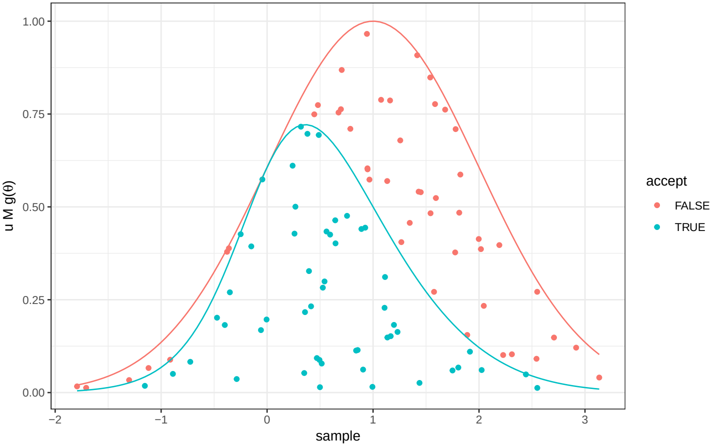
$$g(\theta) = \frac{1}{\sqrt{2\pi}} e^{-(\theta-y)^2/2}$$

with

$$M^\dagger = \sup_\theta \frac{q(\theta|y)}{g(\theta)} = \sup_\theta \frac{e^{-(y-\theta)^2/2} \frac{1}{(1+\theta^2)}}{\frac{1}{\sqrt{2\pi}} e^{-(\theta-y)^2/2}} = \frac{\sqrt{2\pi}}{(1+\theta^2)} \le \sqrt{2\pi}$$

The acceptance rate is $1/M = c(y)/M^\dagger = 1.3056085/\sqrt{2\pi} = 0.5208624$.

Example: Normal-Cauchy model




```
Observed acceptance rate was 0.52
```

Heavy-tailed proposals

Suppose our target is a standard Cauchy and our (proposed) proposal is a standard normal, then

$$\frac{p(\theta|y)}{g(\theta)} = \frac{\frac{1}{\pi(1+\theta^2)}}{\frac{1}{\sqrt{2\pi}}e^{-\theta^2/2}}$$

and

$$\frac{\frac{1}{\pi(1+\theta^2)}}{\frac{1}{\sqrt{2\pi}}e^{-\theta^2/2}} \stackrel{\theta\to\infty}{\longrightarrow} \infty$$

since e^{-a} converges to zero faster than $1/(1+a)$. Thus, there is no value M such that $M\,g(\theta) \ge p(\theta|y)$ for all θ.

TL;DR the condition $M\,g(\theta) \ge p(\theta|y)$ requires the proposal to have tails at least as thick (heavy) as the target.