

Why simulations?

Frequentist statistics (θ fixed, x (and $\hat{\theta}$) random)

Properties of MLEs $\hat{\theta} = \arg \max_{\theta} L(\theta; x)$ Included in Part 3
such as $E(\hat{\theta})$, $\text{Var}(\hat{\theta})$ and its distribution often impossible to obtain analytically.

Instead simulate many realizations $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_n^*$ from an estimated model and estimate $E(\hat{\theta})$ by

$$\widehat{E(\hat{\theta})} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_i^*$$

Bayesian statistics (x fixed, θ random) Included in Part 2

Joint posterior density

$$\pi(\theta | x) \propto L(\theta; x) \pi(\theta)$$

difficult to deal with analytically.

Instead sample from $\pi(\theta | x)$ using MCMC.

Sanity checking: Analytic results $\longleftrightarrow$ Simulations

Part 1: The basics

Methods for simulating from common continuous and discrete distributions.

Pseudorandom number generators

von Neumann (1946). Square X_n and take 4 middle digits as the next "random" number X_{n+1}

Ex. $X_n = 1112$

$$X_n^2 = 01236544$$

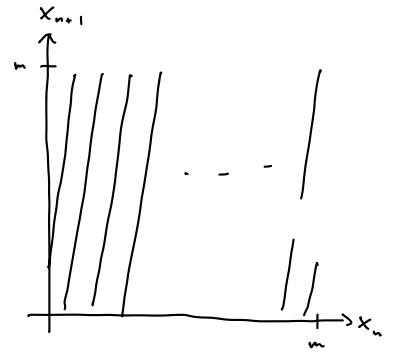
$\underbrace{\quad\quad\quad}_{X_{n+1}}$

Linear congruential generators (e.g. Lehmer 1951, glibc, gcc)

$$X_{n+1} = (aX_n + c) \bmod m$$

Sinclair ZX81 computer (1981)

$$X_{n+1} = (75X_n + 74) \bmod (2^{16} + 1)$$



R demo

R default: Mersenne-Twister (1998)
Period of $2^{19937} - 1$

To change:

`> RNG(kind = "...")`

To repeat same sequence many times
(e.g. for reproducibility):

`> set.seed(7)`

Inversion method (Probability integral transform (PIT)-method)

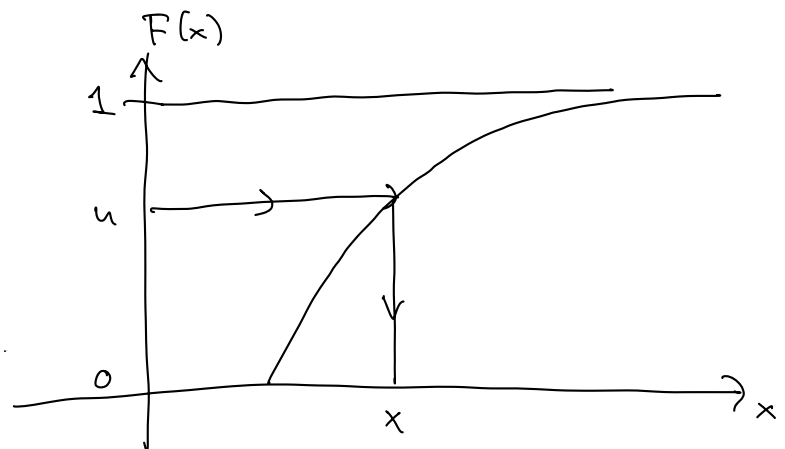
Aim: Simulate $X \sim F_{\theta}(x)$ given $U \sim \text{unif}(0, 1)$
where $F_{\theta}(x) = P(X \leq x)$ is the cdf of X .

Alg.:

$$u \sim \text{Unif}(0, 1)$$

$$x = F_{\theta}^{-1}(u)$$

return x



Proof: The cdf of X becomes

$$F_X(x) = P(X \leq x)$$

$$= P(F^{-1}(U) \leq x)$$

$$= P(U \leq F(x)) \text{ if } F \text{ has an inverse } F^{-1}$$

$$= F_U(F(x))$$

$$= F(x)$$

Ex.: If $X \sim \exp(\lambda)$, then $f(x) = \lambda e^{-\lambda x}$, $F(x) = 1 - e^{-\lambda x}$

If

$$x = F^{-1}(u)$$

then

$$u = F(x) = 1 - e^{-\lambda x}$$

$$\ln(1-u) = -\lambda x$$

$$x = -\frac{1}{\lambda} \ln(1-u) = F^{-1}(u)$$

R dem.

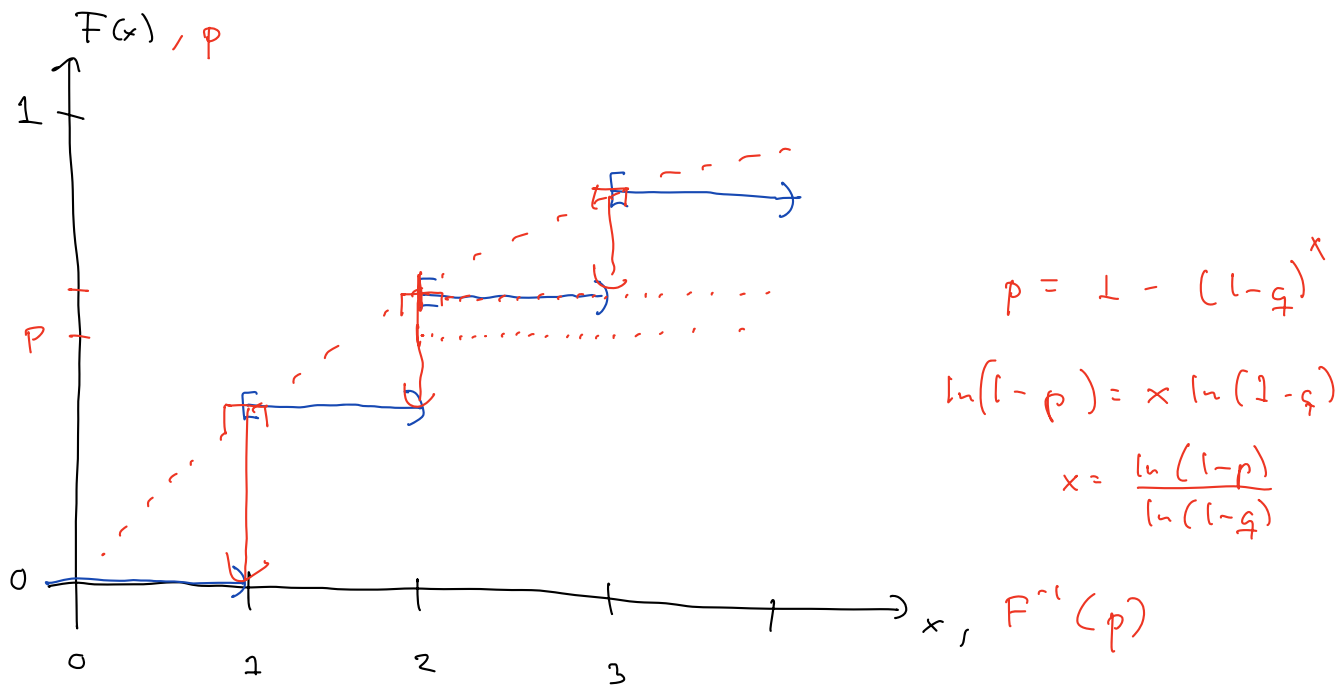
For continuous random variables F^{-1} is known as the quantile function.

Generally (including for discrete random variables) we usually define the quantile function as

$$F^{-1}(p) = \inf \{x \in \mathbb{R} : p \leq F(x)\} \quad (1)$$

(convention followed by g... - functions in $\mathbb{R}$)

Ex.: If $X \sim \text{geom}(q)$ then the pmf is $p(x) = (1-q)^{x-1}q$
 for $x=1, 2, \dots$ and $F(x) = P(X \leq x) = 1 - P(X > x)$
 $= 1 - P(\text{No failures in first } x \text{ trials}) = 1 - (1-q)^x$ (right continuous)



$$F^{-1}(p) = \inf \{x \in \mathbb{R} : p \leq F(x)\} = \dots = \left\lceil \frac{\ln(1-p)}{\ln(1-q)} \right\rceil \quad (\text{left continuous})$$

Alg: $u \sim \text{Unif}(0, 1)$

$$x = \left\lceil \frac{\ln(1-u)}{\ln(1-q)} \right\rceil$$

The quantile function

F^{-1} defined by (1) almost surely a left inverse of F
 in that $F^{-1}(F(X)) = X$ almost surely.

(but $F(F^{-1}(U)) \neq U$ almost surely)

Inversion method only using pmf

$$p(x) = p_i \text{ for } x = x_i, \quad i = 1, 2, \dots, k, \quad x_1 < x_2 < \dots < x_k$$

Alg.

$$u \sim \text{Unif}(0, 1)$$

$$\bar{i} \leftarrow 1$$

$$F \leftarrow p_1$$

$$\text{while } u > F$$

$$\bar{i} \leftarrow \bar{i} + 1$$

$$F \leftarrow F + p_{\bar{i}}$$

Find i
such that

$$F(x_{i-1}) < u \leq F(x_i)$$

end while

return $x_{\bar{i}}$

$$\text{Ex: } X \sim \text{Bern}(p)$$

$$u \sim \text{Unif}(0, 1)$$

$$\text{if } (u < p)$$

$$x \leftarrow 1$$

else

$$x \leftarrow 0$$

return x

Simulating the generating process:

Ex.: $X \sim \text{negbin}(r, p)$

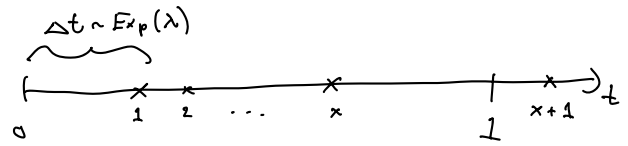
FFFS FFFS F S F ... FFS
1 2 3 ... r

```
s ← 0
x ← 0
while (s < r)
  u ~ Unif(0, 1)
  x ← x + 1
  if (u < p)
    s ← s + 1
  end if
end while
return x
```

Ex.: $X \sim \text{bin}(n, p)$

```
x ← 0
for (i = 1, 2, ..., n)
  u ~ Unif(0, 1)
  if (u < p)
    x ← x + 1
  end if
end for
return x
```

Ex.: $X \sim \text{Poisson}(\lambda)$

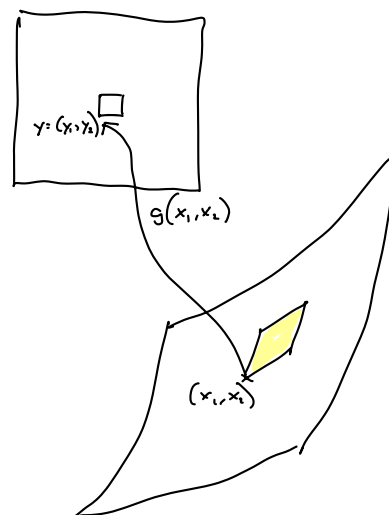


```
x ← 0
t ← 0
do
  Δt ~ Exp(λ)
  t ← t + Δt
  if (t > 1)
    break
  x ← x + 1
end do
return x
```

Transformation formula for joint densities

Suppose that $X = (X_1, X_2)$ has joint pdf $f_X(x_1, x_2)$ and that $Y = (Y_1, Y_2) = g(X_1, X_2)$ is a one-to-one differentiable function with inverse g^{-1} . Then

$$f_Y(y_1, y_2) = f_X(g^{-1}(y_1, y_2)) \underbrace{\left| \begin{array}{cc} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{array} \right|}_{= f_X(x_1, x_2)} \underbrace{|J|}_{|J|}$$



Box-Muller method

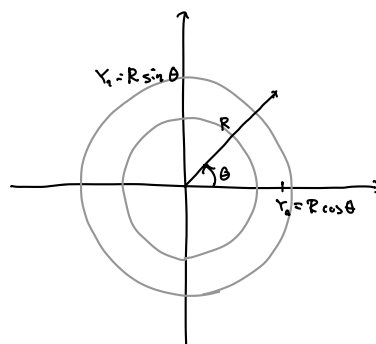
Aim: Simulate from the standard normal

Inversion method? $Y_1 = \Phi^{-1}(U)$?

Consider instead

$Y_1, Y_2 \stackrel{i.i.d.}{\sim} N(0, 1)$. Then $f_Y(y_1, y_2) = \frac{1}{2\pi} e^{-\frac{y_1^2 + y_2^2}{2}}$

$$= \frac{y_1^2 + y_2^2}{2}$$



Joint pdf of corresponding polar coordinates

$$f_{R, \theta}(r, \theta) = f_{Y_1, Y_2}(y_1, y_2) \left| \begin{array}{cc} \frac{\partial y_1}{\partial r} & \frac{\partial y_2}{\partial r} \\ \frac{\partial y_1}{\partial \theta} & \frac{\partial y_2}{\partial \theta} \end{array} \right|$$

$$= \frac{1}{2\pi} e^{-\frac{1}{2}(y_1^2 + y_2^2)} \begin{vmatrix} \cos \theta & \sin \theta \\ -r \sin \theta & r \cos \theta \end{vmatrix}$$

$$= \frac{1}{2\pi} e^{-\frac{r^2}{2}} (r \cos^2 \theta + r \sin^2 \theta)$$

$$= \frac{1}{2\pi} r e^{-\frac{r^2}{2}}$$

$$\underbrace{\quad}_{f_\theta(\theta)} \underbrace{\quad}_{f_R(r)}$$

$$= \frac{r^2}{2}$$

i.e. $\theta \sim \text{Unit}[0, 2\pi]$, and $f_R(r) = r e^{-\frac{r^2}{2}}$.

$$y_1 = r \cos \theta$$

$$y_2 = r \sin \theta$$

R has cdf

$$F_R(r) = \int_0^r x e^{-\frac{x^2}{2}} dx = \left[-e^{-\frac{x^2}{2}} \right]_0^r = 1 - e^{-\frac{r^2}{2}}$$

with inverse

$$u = 1 - e^{-\frac{r^2}{2}}$$

$$\sqrt{-2 \ln(1-u)} = F_R^{-1}(u) = r$$

Alg. (Box-Muller).

$$u \sim \text{Unif}(0,1)$$

$$\theta \sim \text{Unif}(0, 2\pi)$$

$$r \leftarrow \sqrt{-2 \ln u}$$

$$y_1 \leftarrow r \cos \theta$$

$$y_2 \leftarrow r \sin \theta$$

return y_1, y_2

But `rnorm` uses slightly slower inversion method based on minimax rational function approximations of $\Phi^{-1}(\cdot)$ obtained via Remez-algorithm.

Marsaglia's polar method

Alg.: repeat

$$x_1 \sim \text{unif}(-1, 1)$$

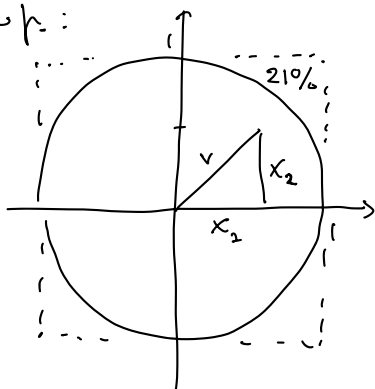
$$x_2 \sim \text{unif}(-1, 1)$$

$$u \leftarrow x_1^2 + x_2^2$$

until $u < 1$

$$\text{return } \sqrt{-\frac{2 \ln u}{u}} (x_1, x_2)$$

Proof:



$$\text{Let } V = x_1^2 + x_2^2$$

$$\begin{aligned} f_{V, \theta}(v, \theta) &= f_{x_1, x_2}(x_1, x_2) \cdot |J| \\ &= \underbrace{\frac{1}{\pi}}_{f_\theta(\theta)} \underbrace{v}_{f_V(v)} \\ &= \frac{1}{2\pi} \cdot 2v \end{aligned}$$

Thus, $F_V(v) = v^2$ and so $U = x_1^2 + x_2^2 = V^2 = F_V(V) \sim \text{Unif}(0, 1)$

From Box-Muller method we know that

$$\begin{aligned} N(0, 1) \stackrel{\text{iid}}{\sim} (X_1, X_2) &= R(\cos \theta, \sin \theta) = \sqrt{-2 \ln U} \left(\frac{X_1}{V}, \frac{X_2}{V} \right) \\ &= \sqrt{-\frac{2 \ln U}{U}} (x_1, x_2) \end{aligned}$$

Ex.: $X \sim \text{Beta}(\alpha, \beta)$. If $U \sim \text{Gamma}(\alpha, 1)$

and $V \sim \text{Gamma}(\beta)$, then $X = \frac{U}{U+V} \sim \text{Beta}(\alpha, \beta)$.

Proof: Let $Y = U+V$. Then

$$f_{X,Y}(x,y) = f_{U,V}(u,v) \begin{vmatrix} \frac{du}{dx} & \frac{du}{dy} \\ \frac{dv}{dx} & \frac{dv}{dy} \end{vmatrix}$$

$$x = \frac{u}{u+v} \quad y = u+v$$

$$= \frac{u}{y}$$

$$u = xy \quad 1-x = \frac{v}{u+v} = \frac{v}{y}$$

$$v = (1-x)y$$

$$= \frac{1}{\Gamma(\alpha)} u^{\alpha-1} e^{-u} \frac{1}{\Gamma(\beta)} v^{\beta-1} e^{-v} \begin{vmatrix} y & x \\ -y & 1-x \end{vmatrix}$$

$$= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1} v^{\beta-1} e^{-\overbrace{(u+v)}^{=y}} [y(1-x) + yx]$$

$$= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} (xy)^{\alpha-1} ((1-x)y)^{\beta-1} y e^{-y}$$

$$= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} y^{\alpha+\beta-1} e^{-y}$$

$$f_X(x) = \int_0^{\infty} f_{X,Y}(x,y) dy = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

Project 1: similar problem for the F-distribution.

(simplified version)

$k f^*(x)$
"

Rejection samp. . Aim: Simulate $X \sim f(x)$ where the support of f is (a, b) and c is an upper bound of f^* .

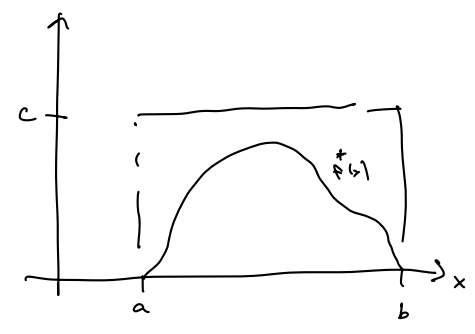
Alg.: repeat

$x \leftarrow \text{Unif}(a, b)$

$u \leftarrow \text{Unif}(0, c)$

until $u < f^*(x)$

return x



Proof: Bayes theorem:

$$f_{X|A}(x) = \frac{P(A|X=x) \frac{1}{b-a}}{P(A)} = \frac{\frac{1}{c} f^*(x) \frac{1}{b-a}}{\int_a^b \frac{1}{c} f^*(x) \frac{1}{b-a} dx} = \frac{f^*(x)}{\int f^*(x) dx} = f(x)$$

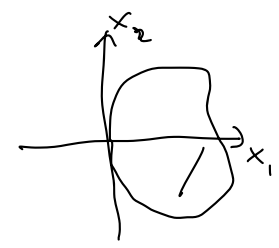
Ratio-of-uniforms method

Idea: Transform problems where support of f^* is unbounded into bivariate density with bounded support.

Alg.: Simulate $(x_1, x_2) \sim \text{Unif}(C)$ where $C = \{(x_1, x_2): 0 \leq x_1 \leq \sqrt{f^*(\frac{x_2}{x_1})}\}$

$y \leftarrow x_2/x_1$

return y



Proof: Joint density of

$$y = \frac{x_2}{x_1}, \quad z = x_1 \Leftrightarrow x_1 = z \quad x_2 = yz$$

becomes

$$f(y, z) = f(x_1, x_2) \begin{vmatrix} \frac{\partial x_1}{\partial y} & \frac{\partial x_1}{\partial z} \\ \frac{\partial x_2}{\partial y} & \frac{\partial x_2}{\partial z} \end{vmatrix}$$

$$= k \cdot \begin{vmatrix} 0 & 1 \\ z & y \end{vmatrix} = kz.$$

Thus, $\sqrt{f^*(y)}$

$$f(y) = \int_0^{\sqrt{f^*(y)}} f(y, z) dz = \left. \frac{kz^2}{2} \right|_0^{\sqrt{f^*(y)}} = \frac{k}{2} f^*(y)$$

If $x^2 f^*(x)$ and $f^*(x)$ are bounded

then $C_1 \subset [0, a] \times [b_-, b_+]$ where

$$a = \sup_x \sqrt{f^*(x)} = \sqrt{\sup_x f^*(x)}, \quad (1)$$

$$b_- = -\sqrt{\sup_{x \leq 0} x^2 f^*(x)}, \quad (2)$$

$$b_+ = \sqrt{\sup_{x \geq 0} x^2 f^*(x)} \quad (3)$$

Proof: $x_1 = z \leq \sqrt{f^*\left(\frac{x_2}{x_1}\right)} \leq \sup_x \left(\sqrt{f^*(x)}\right) = \sqrt{\sup_x f^*(x)} = a$

If $x_2 \geq 0$ then

$$x_1 \leq \sqrt{f^*\left(\frac{x_2}{x_1}\right)}$$

$$x_2 \leq \frac{x_2}{x_1} \sqrt{f^*\left(\frac{x_2}{x_1}\right)} = \sqrt{\left(\frac{x_2}{x_1}\right)^2 f^*\left(\frac{x_2}{x_1}\right)} \leq \sqrt{\sup_{x \geq 0} x^2 f^*(x)} = b_+$$

If $x_2 \leq 0$ then

$$x_2 \geq \frac{x_2}{x_1} \sqrt{f^*\left(\frac{x_2}{x_1}\right)} \geq -\sqrt{\sup_{x \leq 0} x^2 f^*(x)} = b_-$$

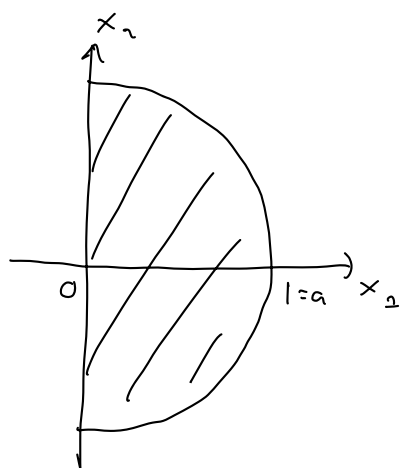
Ex.: $Y \sim \text{Cauchy}(0, 1)$ with $f^*(y) = \frac{1}{1+y^2}$

$$a = \sqrt{\sup_x f^*(x)} = 1$$

$$0 \leq x_1 \leq \sqrt{f^*\left(\frac{x_2}{x_1}\right)} \Rightarrow x_1^2 \leq \frac{1}{1 + \left(\frac{x_2}{x_1}\right)^2} \Rightarrow 1 \leq \frac{1}{x_1^2 + x_2^2}$$

$$\Rightarrow x_1^2 + x_2^2 \leq 1$$

$$\sqrt{1 - x_2^2} \leq x_1 \leq \sqrt{1 + x_2^2}$$



Alg.:

repeat

$$x_2 \sim \text{Unif}(0, 1)$$

$$x_1 \sim \text{Unif}(-1, 1)$$

$$\text{until } x_1^2 + x_2^2 \leq 1$$

$$y \leftarrow x_2 / x_1$$

return x_1

Improvement: If R has density $f(r) = 2r$ for $0 < r < 1$ and $\Theta \sim \text{unif}\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ then $X_1 = R \cos \Theta$, $X_2 = R \sin \Theta$ has a unit density on C .

$$\text{But } Y = \frac{X_2}{X_1} = \frac{R \sin \Theta}{R \cos \Theta} = \tan \Theta \quad (R \text{ cancels})$$

So the alg. simplifies to

$$\Theta \leftarrow \text{unif}\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$$

$$y \leftarrow \tan(\Theta)$$

return y

equivalent to the inversion method.

$$Ex: Y \sim N(0, 1), \quad f^*(y) = e^{-\frac{y^2}{2}}$$

$$a = \sqrt{\sup_x f^*(x)} = 1$$

$$\frac{d}{dx} \left(x^2 e^{-\frac{x^2}{2}} \right) = x^2 (-x) e^{-\frac{x^2}{2}} + 2x e^{-\frac{x^2}{2}} = 0$$

$$2x - x^3 = 0$$

$$x(2 - x^2) = 0$$

$$x = \pm \sqrt{2}$$

$$\sup (x^2 f(x)) = 2 e^{-\frac{2}{2}} = 2 e^{-1}$$

$$b_+ = \sqrt{\sup_x x^2 f(x)} = \sqrt{2} e^{-\frac{1}{2}}$$

$$b_- = -\sqrt{2} e^{-\frac{1}{2}}$$

Alg.: repeat

$$x_1 \sim \text{unif}(0, 1)$$

$$x_2 \sim \text{unif}(-\sqrt{2} e^{-\frac{1}{2}}, \sqrt{2} e^{-\frac{1}{2}})$$

$$\text{until } x_2^2 \leq -4x_1^2 \ln x_1$$

$$y \leftarrow x_2 / x_1$$

return y .

R demo

Ex.: $Y \sim \text{Gamma}(\alpha, \beta)$, $f^*(y) = y^{\alpha-1} e^{-y}$

Mode at

$$\frac{d \ln f^*(x)}{dx} = \frac{d}{dx} \left((\alpha-1) \ln x - x \right) = \frac{\alpha-1}{x} - 1 = 0 \Rightarrow x = \alpha-1$$

if $\alpha \geq 1$. Thus

$$a = \sqrt{\sup_x f^*(x)} = \sqrt{(\alpha-1)^{\alpha-1} e^{-(\alpha-1)}} = \left(\frac{\alpha-1}{e} \right)^{\frac{\alpha-1}{2}}$$

Similarly,

$$\frac{d}{dx} \ln \left(x^2 f^*(x) \right) = \frac{d}{dx} \left[(\alpha+1) \ln x - x \right] = \frac{\alpha+1}{x} - 1 = 0 \Rightarrow x = \alpha+1$$

for all $\alpha > 0$. Thus

$$b_+ = \sqrt{\sup_{x \geq 0} x^2 f^*(x)} = \sqrt{(\alpha+1)^{\alpha+1} e^{-(\alpha+1)}} = \left(\frac{\alpha+1}{e} \right)^{\frac{\alpha+1}{2}}$$

$$b_- = \sqrt{\inf_{x \leq 0} x^2 f^*(x)} = 0$$

Alg.: repeat

$$x_1 \sim \text{Unif}(0, a)$$

$$x_2 \sim \text{Unif}(0, b_+)$$

$$\text{until } x_1 \leq \sqrt{\left(\frac{x_2}{x_1} \right)^{\alpha-1} e^{-\frac{x_2}{x_1}}}$$

$$y \leftarrow x_2 / x_1$$

return y

R demo

Simulation via mixtures

3.1

Many pdfs/pmfs $f(x)$ can be seen as a countable mixture

$$f(x) = \sum_y f(x|y) f(y) = \sum_y f(x, y)$$

or as an uncountable mixture

$$f(x) = \int_{-\infty}^{\infty} f(x|y) f(y) dy = \int_{-\infty}^{\infty} f(x, y) dy$$

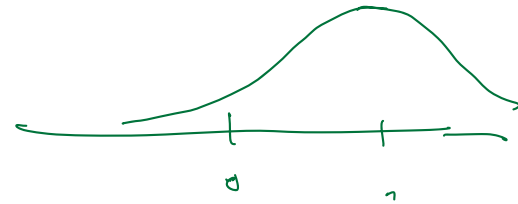
If we can simulate from $f(y)$ and $f(x|y)$ then we can simulate from $f(x)$ using

Alg.: $y \sim f(y)$

$x \sim f(x|y)$

return x

Ex. $f^*(x) = \cosh(x) e^{-\frac{1}{2}(x-1)^2}$



$$= \frac{e^x + e^{-x}}{2} e^{-\frac{1}{2}(x-1)^2}$$

$$= \frac{1}{2} e^{-\frac{1}{2}[x^2 - 4x + 1]} + \frac{1}{2} e^{-\frac{1}{2}[x^2 + 1]}$$

$$= \frac{1}{2} e^{-\frac{1}{2}[x^2 - 4x + 4 - 4 + 1]} + \frac{1}{2} e^{-\frac{1}{2}(x^2 + 1)}$$

$$= \frac{1}{2} e^{\frac{3}{2}} e^{-\frac{1}{2}(x-2)^2} + \frac{1}{2} e^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}$$

$$\propto w \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2} + (1-w) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

$$\frac{w}{\sqrt{2\pi} e^{3/2}} = \frac{1-w}{\sqrt{2\pi} e^{-1/2}}$$

$$w e^{-3/2} = e^{1/2} (1-w)$$

$$w = \frac{e^{1/2}}{e^{3/2} + e^{1/2}} = \frac{1}{1 + e^{-2}} \approx 0.881$$

Alg: $u \sim \text{unif}(0, 1)$
 if $(u < w)$
 $x \sim N(2, 1)$
 else
 $x \sim N(0, 1)$
 return x

Ex: If $\lambda \sim \text{Gamma}(\alpha, \beta)$ and $X|\lambda \sim \text{Pois}(\lambda)$
 then $X \sim \text{neg bin}(\dots, \dots)$.

$$M_\lambda(t) = \left(\frac{1}{1 - \beta t} \right)^\alpha$$

$$M_{X|\lambda}(t) = e^{\lambda(e^t - 1)}$$

$$M_X(t) = E(E e^{tX} | \lambda)$$

$$= E M_{X|\lambda}(t)$$

$$= E e^{\lambda(e^t - 1)}$$

$$= M_\lambda(e^t - 1)$$

$$= \left(\frac{1}{1 - \beta(e^t - 1)} \right)^\alpha$$

Alternatively:

$$f_X(x) = \int_0^\infty f_X(x|\lambda) f_\lambda(\lambda) d\lambda$$

$$= \int_0^\infty \frac{\lambda^x e^{-\lambda}}{x!} \frac{\lambda^{\alpha-1} e^{-\lambda \beta}}{\Gamma(\alpha)} d\lambda$$

= ...

$$= \left(\frac{1}{1 + \beta - \beta e^t} \right)^\alpha$$

$$= \left(\frac{\frac{1}{1+\beta}}{1 - \frac{\beta}{1+\beta} e^t} \right)^\alpha$$

$$= \left(\frac{p}{1 - (1-p)e^t} \right)^\alpha,$$

the mgf of $X \sim \text{neg bin}(\alpha, p)$ where $p = \frac{1}{1+\beta}$.

Alg. to simulate $X \sim \text{neg bin}(\alpha, p)$

$$\beta \leftarrow \frac{1}{p} - 1$$

$$\lambda \sim \text{Gamma}(\alpha, \beta)$$

$$X \sim \text{pois}(\lambda)$$

return X .

Ex.: $T \sim \frac{Z}{\sqrt{V/n}} \sim t_n$. Conditional on $V=v$, T is normal

with $\text{Var}(T|V=v) = \frac{1}{(\sqrt{v/n})^2} = \frac{n}{v}$ and zero mean, i.e.

$$f_{T|V=v}(t) = \frac{1}{\sqrt{2\pi n/v}} e^{-\frac{vt^2}{2n}}$$

Pdf of V

$$f_V(v) = \frac{2^{n/2}}{\Gamma(n/2)} v^{\frac{n}{2}-1} e^{-\frac{v}{2}}$$

Marginal pdf of T is the mixture

$$f_T(t) = \int_0^\infty f_{T|V=v}(t) f_V(v) dv = \dots \propto \left(1 + \frac{t^2}{n}\right)^{-\frac{v+1}{2}}$$

Multivariate normal

$$\underline{x} = (x_1, \dots, x_d) \sim N_d(\underline{\mu}, \underline{\Sigma}) \text{ if } f(\underline{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\underline{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})^T \underline{\Sigma}^{-1}(\underline{x} - \underline{\mu})\right)$$

- Characterization

Any linear transformation $\underline{y} = A\underline{x} + b$ also (multivariate) normal.

- Marginal and conditional distributions

$$\text{Let } Q = \underline{\Sigma}^{-1}, \underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \underline{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \underline{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

$$\text{Then } x_1 \sim N(\mu_1, \Sigma_{11}), \quad x_2 \sim N(\mu_2, \Sigma_{22})$$

and

$$x_1 | x_2 \sim N\left(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}\right)$$

$$\stackrel{(*)}{=} N\left(\mu_1 - Q_{11}^{-1} Q_{12} (x_2 - \mu_2), Q_{11}^{-1}\right)$$

Proof^(*): Suppose w.l.o.g. that $\underline{\mu} = \underline{0}$

$$f_{x_1 | x_2}(x_1) \propto \exp\left(-\frac{1}{2} \begin{pmatrix} x_1^T & x_2^T \end{pmatrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right)$$

$$= \exp\left(-\frac{1}{2} \left(x_1^T Q_{11} x_1 + x_2^T Q_{21} x_1 + x_1^T Q_{12} x_2 + x_2^T Q_{22} x_2 \right)\right)$$

$$\propto \exp\left(-\frac{1}{2} (x_1 - \mu_{1|2})^T Q_{11} (x_1 - \mu_{1|2})\right)$$

$$\propto \exp\left(-\frac{1}{2} \left(x_1^T Q_{11} x_1 - \mu_{1|2}^T Q_{11} x_1 - x_1^T Q_{11} \mu_{1|2} \right)\right)$$

Equating coefficients, $-Q_{11} \mu_{1|2} = Q_{12} x_2$

$$\mu_{1|2} = -Q_{11}^{-1} Q_{12} x_2$$

- Quadratic forms. If Σ has full rank $p \leq d$ (p nonzero eigenvalues), then

$$(\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \sim \chi_p^2$$

$$(\underline{x} - \underline{\mu})^T P D^{-1/2} D^{-1/2} P^T (\underline{x} - \underline{\mu}) \sim N_d(\underline{0}, I_d)$$

- Simulation:

Let $\underline{x} \sim N_d(\underline{0}, I_n)$. Then $\underline{y} = \underline{\mu} + A\underline{x} \sim N_d(\underline{\mu}, AA^T)$

Thus, if $AA^T = \Sigma$, then $\underline{y} \sim N_d(\underline{\mu}, \Sigma)$.

Possible choices of A :

$A = PD^{1/2}$ where $PDP^T = \Sigma$ is the eigen decomposition of Σ

$A = \text{chol}(\Sigma) = L$ (lower triangular matrix). (In R chol returns L^T)

Finding L :

We have

$$LL^T = \begin{bmatrix} L_{11}^2 & & \\ L_{21}L_{11} & L_{21}^2 + L_{22}^2 & \\ L_{31}L_{11} & L_{31}L_{21} + L_{32}L_{22} & L_{31}^2 + L_{32}^2 + L_{33}^2 \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & & \\ \Sigma_{21} & \Sigma_{22} & \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{bmatrix}$$

Given Σ , first solve for L_{11} , then L_{21} , L_{22} , L_{31} , L_{32} , L_{33} , ...

Cost $O(n^3)$

Rejection sampling

Aim: Simulate $X \sim f(x) = kf^*(x)$, only $f^*(x)$ known.

Can simulate from proposal $g(x)$.

Let $c > 0$ such that $f^*(x) \leq cg(x)$ for all $x \in \mathbb{R}$.

Alg.: repeat

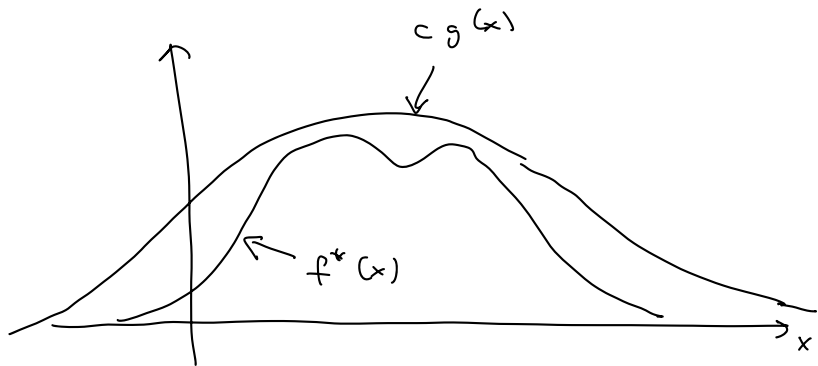
$$x \sim g(x)$$

$$u \sim \text{unif}(0, 1)$$

$$\alpha \leftarrow \frac{1}{c} \frac{f^*(x)}{g(x)}$$

until $u < \alpha$

return x .



Proof:

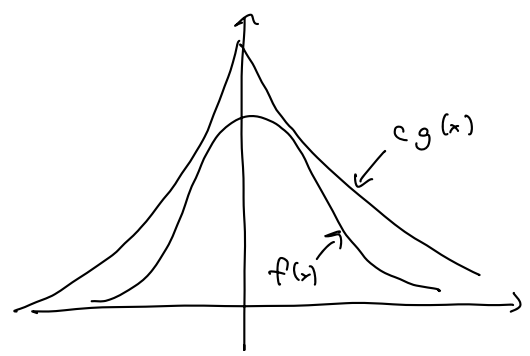
$$\begin{aligned} f_{X|U < \alpha}(x) &= \frac{P(U < \alpha | X=x) g(x)}{P(U < \alpha)} = \frac{P(U < \alpha | X=x) g(x)}{\int P(U < \alpha | X=x) g(x) dx} \\ &= \frac{\frac{1}{c} \frac{f^*(x)}{g(x)} g(x)}{\int \frac{1}{c} \frac{f^*(x)}{g(x)} g(x) dx} \\ &= \frac{f^*(x)}{\int f^*(x) dx} = f(x). \end{aligned}$$

Note that overall acceptance probability

$$P(U < \alpha) = \frac{1}{c} \int f^*(x) dx = \frac{1}{c} \text{ if } f^*(x) = f(x).$$

Ex.: $X \sim N(0,1)$, $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Use $g(x) = \frac{\lambda}{2} e^{-\lambda|x|}$ as proposal.



Optimal choice of c and λ ?
Given λ what is $\sup_x \left(\frac{f(x)}{c g(x)} \right)$?

The acceptance probability (conditional on $X=x$) is

$$\frac{f(x)}{c g(x)} = \frac{2 e^{-x^2/2}}{\sqrt{2\pi} c \lambda e^{-\lambda|x|}} = \frac{1}{\lambda c} \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2} + \lambda x}$$

$$= \frac{1}{\lambda c} \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}[(x-\lambda)^2 - \lambda^2]}$$

and has its maximum of

$$\frac{1}{\lambda c} \sqrt{\frac{2}{\pi}} e^{\frac{\lambda^2}{2}}$$

at $x=\lambda$. We need

$$\frac{1}{\lambda c} \sqrt{\frac{2}{\pi}} e^{\frac{\lambda^2}{2}} \leq 1$$

and thus

$$c \geq \frac{1}{\lambda} \sqrt{\frac{2}{\pi}} e^{\frac{\lambda^2}{2}} = c(\lambda)$$

The overall acceptance probability (known since f is known)

$$P(U < \alpha) = \frac{1}{c(\lambda)} = \sqrt{\frac{\pi}{2}} \lambda e^{-\frac{\lambda^2}{2}} \quad (*)$$

has its maximum when

$$\frac{d}{d\lambda} \left(\ln \lambda - \frac{\lambda^2}{2} \right) = \frac{1}{\lambda} - \lambda = 0 \Rightarrow \lambda = 1$$

for which $c(\lambda) = \sqrt{\frac{2}{\pi}} e^{\frac{1}{2}} = 1.315$ and $P(U < \alpha) = \frac{1}{c} = 0.7602$

Implementation:

repeat

$$u_1 \sim \text{Unif}(0, 1)$$

$$x \leftarrow -\log(u)$$

$$u_2 \leftarrow \text{Unif}(0, 1)$$

$$\text{if } (u_2 < \frac{1}{2})$$

$$x \leftarrow -x$$

$$\alpha \leftarrow e^{-\frac{x^2}{2} + |x| + \frac{1}{2}}$$

$$u_2 \leftarrow \text{Unif}(0, 1)$$

$$\text{until } (u_2 < \alpha)$$

return x

$$\alpha = \frac{f(x)}{c g(x)} = \frac{e^{-\frac{x^2}{2} + |x| + \frac{1}{2}}}{\sqrt{2\pi}} = e^{-\frac{x^2}{2} + |x| + \frac{1}{2}}$$

Note: If only f^* was known we would

need to maximise $(*)$ by experimental tuning of λ .

Rejection sampling multivariate densities

Ex: Sampling uniformly within a unit n -ball.

$$f(\underline{x}) = f(x_1, \dots, x_n) = \frac{1}{V_n} \quad \text{for } \|\underline{x}\|_2 \leq 1$$

$$\text{where } V_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)}$$

Proposal

$$g(\underline{x}) = \frac{1}{2^n} \quad \text{for } -1 \leq x_i \leq 1, \quad i=1, 2, \dots, n.$$

$$f(\underline{x}) = \frac{1}{V_n} \leq c g(\underline{x}) = \frac{c}{2^n} \quad \text{for } c \geq \frac{2^n \Gamma(\frac{n}{2} + 1)}{\pi^{n/2}}$$

$$\text{Accept with prob. } P(U < \alpha | X = \underline{x}) = \alpha = \frac{f(\underline{x})}{c g(\underline{x})} = \begin{cases} 1 & \text{if } \|\underline{x}\|_2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Overall acceptance probability

$$P(U \leq \alpha) = \int_{V_n} \frac{f(\underline{x})}{c g(\underline{x})} \cdot g(\underline{x}) d\underline{x} = \frac{1}{c} = \frac{\pi^{n/2}}{2^n \Gamma(\frac{n}{2} + 1)}$$

n	$U \leq \alpha$
1	1
2	$\pi/4 = 0.78$
3	$\frac{\pi^{3/2}}{2^3 \Gamma(\frac{3}{2} + 1)} = \frac{\pi^{3/2}}{2^3 \cdot \frac{3}{2} \cdot \frac{1}{2} \sqrt{\pi}} = \frac{\pi}{2 \cdot 3} = 0.52$
4	$\frac{\pi^2}{2^4 \Gamma(3)} = \frac{\pi^2}{2^4 \cdot 2 \cdot 1} = 0.308$
10	0.0025

Rejection sampling cont.

3.2

Difficulties

Alternative method

Finding c

Weighted resampling

Finding good proposal

Adaptive rejection sampling

Alias method (x discrete, finite support)

Weighted resampling

Aim: Sample from $f^*(x)$.

Alg.:

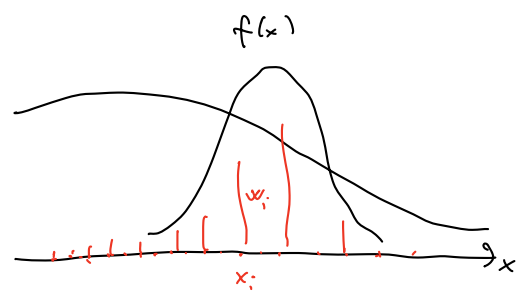
$$x_1, \dots, x_n \stackrel{iid}{\sim} g(x)$$

for $i = 1, \dots, n$

$$w_i \leftarrow \frac{f(x_i)/g(x_i)}{\sum_{i=1}^n f(x_i)/g(x_i)}$$

$$Y_1, \dots, Y_m \stackrel{iid}{\sim} f_x(y) \text{ where } f_x(y) = \begin{cases} w_i & \text{for } y = x_i, i = 1, 2, \dots, n \\ 0 & \text{elsewhere} \end{cases}$$

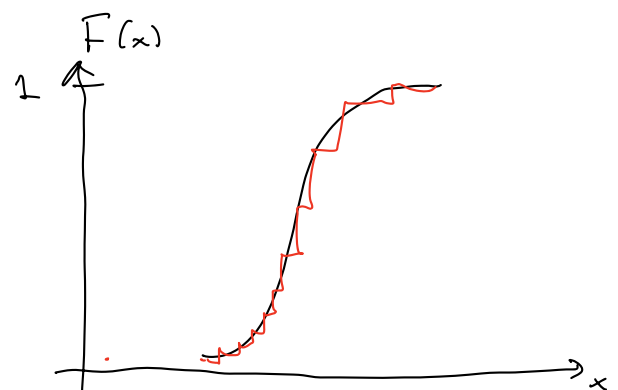
return $Y_1, \dots, Y_m$



The discrete distribution f_x approximates f in the sense that its cdf

$$F_x(a) = P(X \leq a)$$

$$= \sum_{i=1}^n w_i \mathbb{I}(x_i \leq a)$$



Alias method (Walker 1974, 1977)

Aim: Simulate X having pmf $P(X=i)=p_i$, $i=1, 2, \dots, n$

Alg.:

Precompute (U_i, K_i) , $i=1, 2, \dots, n$

$U_i \leftarrow np_i$, $K_i \leftarrow NA$

While $U_i > 1$, $U_j < 1$, $K_j = NA$ for $i \neq j$

$U_i \leftarrow U_i - (1 - U_j)$

$K_j \leftarrow i$

Operation

$u \sim \text{unif}(0, 1)$

$i \leftarrow \lfloor nu \rfloor + 1$

$y \leftarrow nu - \lfloor nu \rfloor$

if $(y < U_i)$

$x \leftarrow i$

else

$x \leftarrow K_i$

return x

Ex.: $n=5$

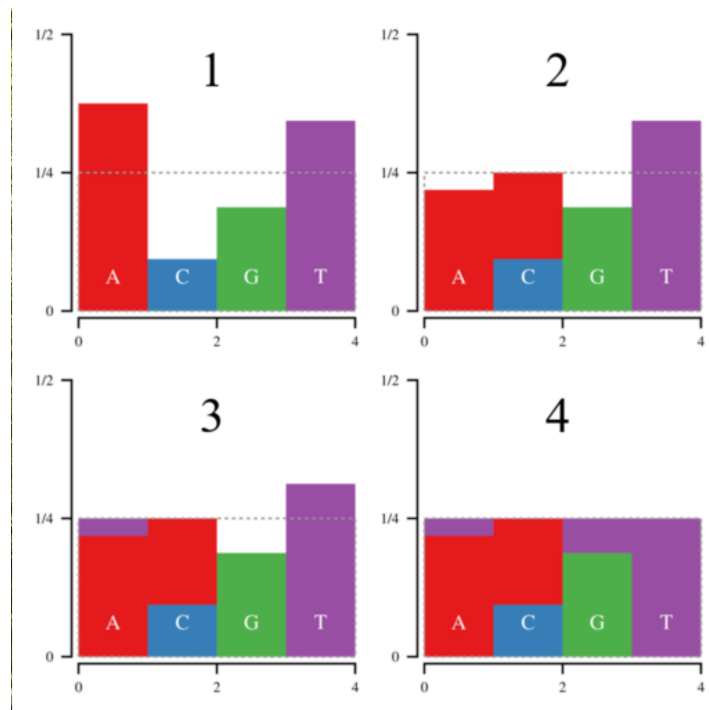
$p = (p_1, \dots, p_5)$

$= (.25, .3, .1, .2, .15)$

$O(n \ln n)$

i	U_i	K_i
1	1.25 0.75	NA 2
2	1.5 1.25 1	NA
3	0.5	NA 1
4	1	NA
5	0.75	NA 2

$O(1)$



Adaptive rejection sample

Assuming that $\ln f(x)$ is concave:

- Construct a piecewise linear log proposal $\ln(cg(x))$ from tangents of $\ln f(x)$ at initial grid points $x_1, x_2, \dots, x_n$.

do

$x \sim g(x)$

(a countable finite
mixture of truncated
exponentials)

$$\alpha \leftarrow \frac{f(x)}{cg(x)}$$

$u \sim \text{unif}(0, 1)$

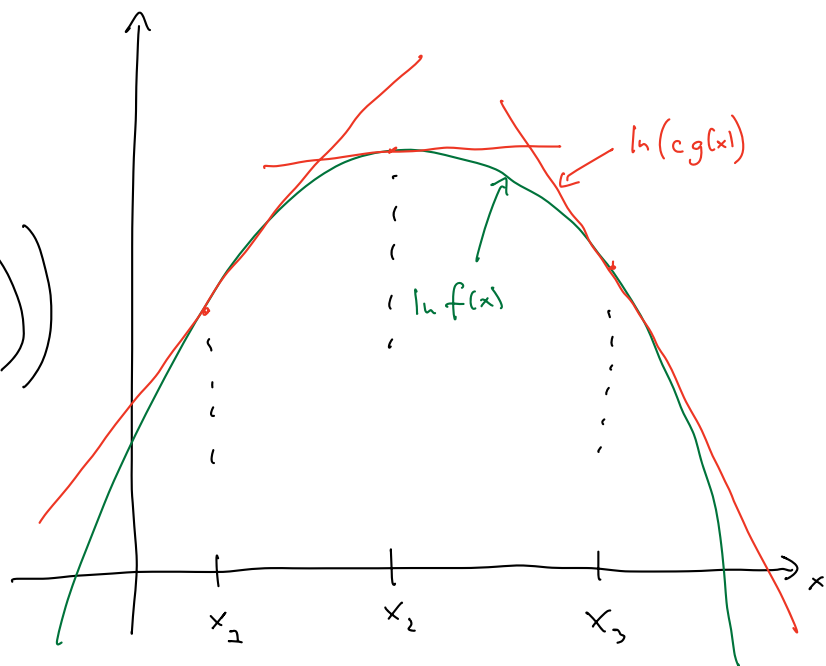
if ($u < \alpha$)

break

improve $\ln(cg(x))$ by adding x to grid

end do

return x



R-package : *ars*

Monte Carlo integration

Aim: Estimate $E(h(X))$ where $X \sim f(x)$.

$$E_X: \underbrace{E(X)}_{h(X)=X}, \quad \text{Var } X = E(\underbrace{(X-\mu)^2}_{h(X)}), \quad P(X > a) = E(\underbrace{\mathbb{I}(X > a)}_{h(X)})$$

Simulate $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} f(x)$ and estimate $E(h(X))$ by

$$\widehat{E(h(X))} = \frac{1}{n} \sum_{i=1}^n h(X_i)$$

Unbiased since

$$E\left(\frac{1}{n} \sum_{i=1}^n h(X_i)\right) = \frac{1}{n} \sum_{i=1}^n E h(X_i) = E h(X)$$

By strong law of large numbers

$$\frac{1}{n} \sum_{i=1}^n h(X_i) \xrightarrow{\text{a.s.}} E h(X)$$

as $n \rightarrow \infty$

Importance sampling

Aim: Estimate $E(h(X))$, $X \sim f(x)$

Alg:

Sample $x_1, x_2, \dots, x_n \stackrel{i.i.d.}{\sim} g(x)$.

Compute

$$\widehat{E(h(X))} = \frac{1}{n} \sum_{i=1}^n h(x_i) \frac{f(x_i)}{g(x_i)}$$

This estimator is unbiased since

$$\begin{aligned} E\left(\widehat{E(h(X))}\right) &= \frac{1}{n} \sum_{i=1}^n E\left(h(X_i) \frac{f(X_i)}{g(X_i)}\right) \\ &= \int_{-\infty}^{\infty} h(x) \frac{f(x)}{g(x)} g(x) dx \\ &= \int_{-\infty}^{\infty} h(x) f(x) dx = E(h(X)). \end{aligned}$$

By the law of large numbers $\widehat{E(h(X))} \xrightarrow[n \rightarrow \infty]{a.s.} E(h(X))$

Optimal choice of $g(x)$ is

$$g(x) = \frac{h(x) f(x)}{\int h(x) f(x) dx}$$

since this leads to

$$\text{Var}\left(\frac{h(X) f(X)}{g(X)}\right) = \text{Var}\left(\frac{h(X) f(X)}{\frac{h(X) f(X)}{\int h(x) f(x) dx}}\right) = \text{Var}\left(\int h(x) f(x) dx\right) = 0$$

In practice choose $g(x)$ such that it is approximately proportional to $h(x)f(x)$.

R demo

Importance sampling est. (self-normalizing version):

4.1

Given only $f^*(x)$ (unnormalized density of X), the alternative estimator

$$\widetilde{E(h(X))} = \frac{\sum_{i=1}^n h(x_i) \frac{f^*(x_i)}{g(x_i)}}{\sum_{i=1}^n \frac{f^*(x_i)}{g(x_i)}}$$

is not unbiased (since $E(\frac{U}{V}) \neq \frac{E(U)}{E(V)}$). But, as $n \rightarrow \infty$,

$$\sum_{i=1}^n \frac{f^*(x_i)}{g(x_i)} \xrightarrow{P} E\left(\frac{f^*(X)}{g(X)}\right) = \int \frac{f^*(x)}{g(x)} g(x) dx = \int f^*(x) dx$$

and

$$\sum_{i=1}^n h(X_i) \frac{f^*(x_i)}{g(x_i)} \xrightarrow{P} E\left(h(X) \frac{f^*(X)}{g(X)}\right) = \dots = \int h(x) f^*(x) dx$$

Thus, using Slutsky's theorem,

$$\widetilde{E(h(X))} \xrightarrow{P} \frac{\int h(x) f^*(x) dx}{\int f^*(x) dx} = \int h(x) f(x) dx = E(h(X)),$$

i.e. $\widetilde{E(h(X))}$ is consistent.

Antithetic sampling

Aim: Reduce variance of Monte Carlo (or importance sampling) estimate $\frac{1}{n} \sum_{i=1}^n h(X_i)$ by making terms negatively correlated.

Generate $U_1, U_2, \dots, U_n \stackrel{\text{i.i.d.}}{\sim} \text{Unit}(0,1)$ and compute

$$\widehat{E(h(X))} = \frac{1}{2n} \left(\sum_{i=1}^n \underbrace{h(F^{-1}(U_i))}_{X_i} + \sum_{i=1}^n \underbrace{h(F^{-1}(1-U_i))}_{X_i^*} \right) = \frac{1}{n} \sum_{i=1}^n \underbrace{\frac{1}{2} (h(F^{-1}(U_i)) + h(F^{-1}(1-U_i)))}_{\text{i.i.d.}} \quad (1)$$

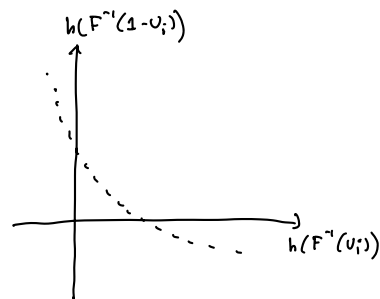
It follows that $h(F^{-1}(U_i)) \sim h(F^{-1}(1-U_i))$

Hence, (1) remains unbiased and converges a.s. to $Eh(X)$.

Also, at least if h (and hence $h \circ F^{-1}$) is mon-tonic,

$$\text{Cov} \left(h(F^{-1}(U_i)), h(F^{-1}(1-U_i)) \right) \leq 0.$$

(see GH, p. 187-188 for a proof)



Hence,

$$\begin{aligned} \text{Var} \left(\widehat{E(h(X))} \right) &= \frac{1}{2n} \text{Var} h(X_i) + \frac{1}{(2n)^2} \cdot n^2 \text{Cov} (h(X_i), h(X_i^*)) \\ &= \frac{1}{2n} \text{Var} h(X_i) (1 + \rho) \\ &\leq \frac{1}{2n} \text{Var} h(X_i), \end{aligned}$$

the variance of an ordinary Monte Carlo estimate based on sample of size $2n$, if $\rho = \text{corr}(h(X_i), h(X_i^*)) \leq 0$

Ex. 1: Find $E(\sqrt{X})$ where $X \sim \text{exp}(1)$.

R demo

Ex. 2. Find $E(h(X))$ where $X \sim N(0,1)$, $h(X) = \frac{X}{2^X - 1}$.

Letting $X_i^* = -X_i$, $X_i^* \sim X_i$ (inversion method not needed)

R demo

Similar ideas can be in conjunction with Markov chain Monte Carlo (MCMC part 2)

Bayesian inference

Recall: From def. of cond. prob.

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B)} \stackrel{\text{L.T.P.}}{=} \frac{P(B|A_i)P(A_i)}{\sum P(B|A_i)P(A_i)} \propto P(B|A_i)P(A_i)$$

for $i=1, 2, \dots, n$ where $A_1, \dots, A_n$ is a partition of sample space B .

For discrete random variables,

$$P(X=x|Y=y) = \frac{\overbrace{P(X=x \cap Y=y)}^{\text{joint pmf of } X, Y}}{\underbrace{P(Y=y)}_{\text{marginal pmf of } Y}} = \frac{P(Y=y|X=x)P(X=x)}{P(Y=y)}$$

For continuous random variables, conditional pdf of X given $Y=y$

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)}$$

or

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}$$

Probabilities are extended to include subjective (personal) degree of belief in different values of unknown parameters (fixed constant) and updated via Bayes theorem

$$\underbrace{f(\theta|x)}_{\text{posterior distr.}} \propto \underbrace{f(x|\theta)}_{\text{likelihood}} \underbrace{f(\theta)}_{\text{prior distr.}}$$

Ex.: $X \sim \text{bin}(n, p)$, n known, p unknown. (e.g. proportion of conservative voters in political survey)

Prior beliefs about p modelled by assuming that $p \sim \text{Beta}(\alpha, \beta)$

Thus, the prior (density) is

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \propto p^{\alpha-1} (1-p)^{\beta-1} \text{ for } 0 \leq p \leq 1$$

The likelihood $L(\theta; x) = f(x|\theta) = \binom{n}{x} p^x (1-p)^{n-x}$

Given the data $X=x$, the posterior density

$$f(p|x) \propto f(x|p)f(p)$$

$$\propto p^x (1-p)^{n-x} p^{\alpha-1} (1-p)^{\beta-1}$$

$$= p^{\alpha+x-1} (1-p)^{\beta+n-x-1}$$

i.e. $p|X=x \sim \text{Beta}(\alpha+x, \beta+n-x)$.

For $n=10, x=0$ and $\alpha=\beta=1$, $p \sim \text{unif}(0,1)$ and

$$p|X=0 \sim \text{Beta}(1, 11)$$

with posterior expectation and standard deviation

$$E(p|X=0) = \frac{1}{1+11} = \frac{1}{12} = 0.083, \quad SD(p|X=0) = \sqrt{\frac{1}{12} \cdot \frac{11}{12} \cdot \frac{1}{13}} = 0.076.$$

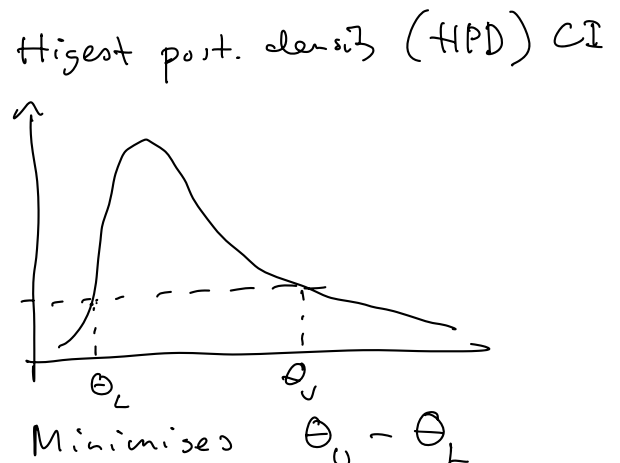
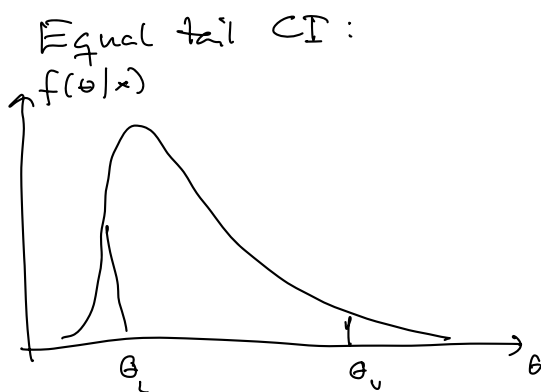
In contrast, MLE of p and MLE of its SE is

$$\hat{p} = \frac{x}{n} = 0, \quad \widehat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.$$

Other summaries of $f(p|x)$, Bayesian point estimates:

- The mode $\text{Mod}(\theta|x) = \arg\max_{\theta} f(\theta|x)$
- The median
- Credible intervals $(\theta_L(x), \theta_U(x))$ s.t.

$$P(\theta_L(x) < \theta < \theta_U(x) | X=x) = 1 - \alpha$$



In contrast, a confidence interval $(\hat{\theta}_L(x), \hat{\theta}_U(x))$ is defined s.t.

$$P(\hat{\theta}_L(x) < \theta < \hat{\theta}_U(x) | \theta) = 1 - \alpha$$

Note that

$$\int \underbrace{P(\theta_L(x) < \theta < \theta_U(x) | X=x)}_{= 1 - \alpha \text{ by construction}} f(x) dx = (1 - \alpha) \int f(x) dx = 1 - \alpha$$

$$\stackrel{LTP}{=} P(\theta_L(x) < \theta < \theta_U(x)) \stackrel{LTP}{=} \int \underbrace{P(\theta_L < \theta < \theta_U | \theta)}_{\text{freq. coverage}} f(\theta) d\theta$$

i.e. frequentist coverage of Bayesian credible interval is $1 - \alpha$ on average across different values of θ weighted by their prior $f(\theta)$. provided that $f(\theta)$ is proper.

Sequential updating

Step 1: Condition on x_1

$$\underbrace{f(\theta | x_1)}_{\text{posterior}} \propto \underbrace{f(x_1 | \theta)}_{\text{prior}} f(\theta)$$

Step 2: Condition also on x_2

$$\underbrace{f(\theta | x_2, x_1)}_{\text{new posterior}} \propto \underbrace{f(x_2 | \theta, x_1)}_{= f(x_2 | \theta) \text{ if } x_2, x_1 | \theta \text{ are indep.}} \underbrace{f(\theta | x_1)}_{\text{old posterior} = \text{new prior}}$$

Same result as

$$\begin{aligned} f(\theta | x_2, x_1) &\propto f(x_2, x_1 | \theta) f(\theta) \\ &= f(x_2 | x_1, \theta) f(x_1 | \theta) f(\theta) \end{aligned}$$

Choice of prior:

Should represent prior beliefs.

- Conjugate priors

A family G of distributions is conjugate with respect to a likelihood $L(\theta; x) = f(x|\theta)$ if both prior $f(\theta)$ and posterior $f(\theta|x)$ is in G .

Ex. 1: $L(p; x) = \binom{n}{x} p^x (1-p)^{n-x}$, $p \sim \text{Beta}(\alpha, \beta)$ is conjugate to $L(p; x)$ since $p|x \sim \text{Beta}(\alpha+x, \beta+n-x)$ if $p \sim \text{Beta}(\alpha, \beta)$.

Note: Simplifies sequential updating.

Ex. 2: $X \sim N(\mu, \sigma^2)$, σ^2 known, μ unknown
$$-\frac{1}{2\sigma^2} (x-\mu)^2$$

$$L(\mu; x) = f(x|\mu) \propto e$$

Conjugate prior is $\mu \sim N(\mu_0, \sigma_0^2)$

$$f(\mu) \propto e^{-\frac{1}{2\sigma_0^2} (\mu-\mu_0)^2}$$

since the posterior is then

$$f(\mu|x) \propto f(x|\mu) f(\mu)$$

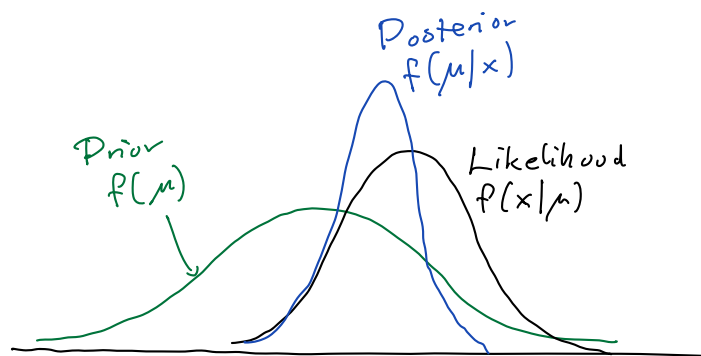
$$\begin{aligned} & -\frac{1}{2} \left[\frac{1}{\sigma^2} (x-\mu)^2 + \frac{1}{\sigma_0^2} (\mu-\mu_0)^2 \right] \\ & = e \\ & -\frac{1}{2} \left[\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 - 2 \left(\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \mu + \dots \right] \\ & \propto e^{-\frac{1}{2} \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \left[\mu - \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}} \right]^2 + \dots} \\ & \propto e \end{aligned}$$

(completing the square)

i.e. $\mu|x \sim N(\mu_1, \sigma_1^2)$ where $\frac{1}{\sigma_1^2} = \underbrace{\frac{1}{\sigma^2}}_{\text{posterior precision}} = \underbrace{\frac{1}{\sigma^2}}_{\text{likelihood precision}} + \underbrace{\frac{1}{\sigma_0^2}}_{\text{prior precision}}$ (add precisions)

and $\mu_1 = \frac{\frac{1}{\sigma^2}x + \frac{1}{\sigma_0^2}\mu}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}$ (average of x and μ weighted by respective precisions)

(same family G)



Ex. 3: $x|\sigma^2 \sim N(\mu, \sigma^2)$, μ known, σ^2 unknown

$$-\frac{1}{2} - \frac{(x-\mu)^2}{2\sigma^2}$$

$$L(\sigma^2; x) \propto (\sigma^2)^{-\frac{1}{2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Reparameterizing in terms of precision $\tau = \frac{1}{\sigma^2}$

$$L(\tau; x) \propto \tau^{\frac{1}{2}} e^{-\frac{(x-\mu)^2}{2}\tau}$$

Conjugate prior is $\tau \sim \text{Gamma}(\alpha, \beta)$

$$f(\tau) \propto \tau^{\alpha-1} e^{-\beta\tau}$$

since

$$f(\tau|x) \propto \tau^{\alpha+\frac{1}{2}-1} e^{-(\beta+\frac{(x-\mu)^2}{2})\tau}$$

i.e.

$$\tau|x \sim \text{Gamma}\left(\alpha+\frac{1}{2}, \beta+\frac{(x-\mu)^2}{2}\right)$$

Accordingly, conjugate family for $\sigma^2 = \frac{1}{\tau}$ is the inverse gamma

Ex. 4-6:

Likelihood	Conjugate
Geom	Beta
Poisson	Gamma
Exp	Gamma

Exponential family likelihoods always have a conjugate prior (GL 2.3.1.)

Improper priors.

Attempt at representing complete lack of prior knowledge about θ

("objectivity") via pdf $f(\theta)$, $\int_{-\infty}^{\infty} f(\theta) d\theta = \infty$, e.g. $f(\theta) = 1$ for $-\infty < \theta < \infty$

Ex. 2 cont.: $X|\mu \sim N(\mu, \sigma^2)$, σ^2 known,

$$\mu \sim N(\mu_0, \sigma_0^2) \quad (1a)$$

prior for μ .

As $\sigma_0^2 \rightarrow \infty$, the posterior

$$f(\mu|x) \rightarrow \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(\mu-x)^2}{2\sigma^2}} \quad (1b)$$

If we instead use the improper prior

$$f(\mu) \propto 1 \quad (2a)$$

as prior for μ , and assume that

$$f(\mu|x) \propto f(x|\mu)f(\mu) \quad (2b)$$

still holds, we obtain

$$f(\mu|x) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot 1 \quad (2c)$$

i.e. the same result as limiting posterior (1b) above.

But note that

$$\lim_{\sigma_0^2 \rightarrow \infty} f(\mu; \mu_0, \sigma_0^2) \rightarrow 0$$

which differs from $f(\mu)$ in (2a).

If a sequence of proper priors converges q -vaguely
(see Bioche & Druilhet, 2016, for def.) to an improper prior $f(\theta)$,
then the limiting posterior (as in 1b)

can be obtained instead via $f(\theta|x) \propto f(x|\theta)f(\theta)$ (as in 2c)
where $f(\theta)$ is the q -vague limiting improper prior.

Some improper priors leads to credible intervals with exact frequentist coverage.
(matching priors)

Non-informative "flat" priors

Aim: "Objectivity".

Ex.: $X \sim \text{bin}(n, p)$, Prior on p : $f(p) = 1$ for $0 < p < 1$.

Criticism: Implies that

$$\theta = \ln \frac{p}{1-p} \quad p = \frac{e^\theta}{1+e^\theta}$$

has a non-uniform density

$$f_\theta(\theta) = f(p) \left| \frac{dp}{d\theta} \right| = \frac{e}{(1+e^\theta)^2} \quad \text{for } -\infty < \theta < \infty,$$

i.e. $\theta \sim \text{logistic}(0, 1)$. So in terms of θ the prior is not "non-informative".

An alternative improper prior is $f(\theta) \propto 1$ for $-\infty < \theta < \infty$ which leads to

$$f(p) = f(\theta) \frac{d\theta}{dp} = \frac{1}{p} - \frac{1}{1-p} = \frac{1}{p(1-p)} = p^{-1}(1-p)^{-1}$$

(the Haldane prior) which can be seen as a (q-vague) limit of a $\text{Beta}(\alpha, \beta)$ as $\alpha, \beta \rightarrow 0$. For the $X|p \sim \text{bin}(n, p)$ likelihood this leads to

and thus $p|X=x \sim \text{Beta}(x, n-x)$

$$E(p|X=x) = \frac{x}{x + n-x} = \frac{x}{n}$$

for $0 < x < n$. For $x=0$ the posterior

$$f(p|x) \propto p^{-1} (1-p)^{n-1}$$

is improper (and similarly for $x=n$).

Jeffreys prior

Aim: Invariance w.r.t. reparameterization

Def.: $f(\theta) \propto |J(\theta)|^{\frac{1}{2}}$
where $J(\theta) = E\left(-\frac{\partial^2}{\partial \theta \partial \theta^T} \ln L(\theta; X)\right) = E\left(\frac{\partial}{\partial \theta} \ln L(\theta; X) \frac{\partial}{\partial \theta^T} \ln L(\theta; X)\right)$
is the Fisher information.
alternative equivalent def.

Proof of invariance (1-dimensional case):

The log-likelihood after reparameterization to $y = h(\theta)$ where h is one-to-one and differentiable is

$$l(y; X) = l_{\theta}(\theta(y); X) \quad \text{where} \quad l_{\theta}(\theta; X) = \ln L(\theta; X).$$

Thus

$$\frac{d}{dy} l(y; X) = \frac{d}{d\theta} l_{\theta}(\theta(y); X) \frac{d\theta}{dy}$$

$$\frac{d^2}{dy^2} l(y; X) = \frac{d^2}{d\theta^2} l_{\theta}(\theta(y); X) \left(\frac{d\theta}{dy}\right)^2 + \frac{d}{d\theta} l_{\theta}(\theta(y); X) \frac{d^2 \theta}{dy^2} \quad (*)$$

and

$$J_y(y) = -E\left(\frac{\partial^2}{\partial y^2} l(y; X)\right) \\ = J_{\theta}(\theta) \left(\frac{d\theta}{dy}\right)^2$$

since the score function (second term in $(*)$) has zero expectation.

Hence, prior based on $J_y(y)$ for the reparameterized model

$$f_y(y) = \sqrt{J_{\theta}(\theta)} \left| \frac{d\theta}{dy} \right| = f_{\theta}(\theta) \left| \frac{d\theta}{dy} \right|$$

is equal to the prior implied by the transformation formula and $J_{\theta}(\theta)$ in first parameterization.

Proof in p -dimensional case

$$\gamma = g(\theta), \quad g: \mathbb{R}^p \rightarrow \mathbb{R}^p$$

$$\underbrace{\frac{\partial}{\partial \gamma}}_{p \times 1} \ell_g(\gamma; X) = \frac{\partial}{\partial \gamma} \ell_\theta(g(\gamma); X) = \underbrace{\frac{\partial \theta}{\partial \gamma}}_{p \times p} \underbrace{\frac{\partial \ell_\theta}{\partial \theta}}_{p \times 1}$$

$$\begin{aligned} J_\gamma(\gamma) &= E \left(\frac{\partial \ell_\gamma}{\partial \gamma} \cdot \frac{\partial \ell_\gamma}{\partial \gamma}^\top \right) = E \left(\underbrace{\frac{\partial \theta}{\partial \gamma}}_{p \times p} \cdot \overbrace{\frac{\partial \ell_\theta}{\partial \theta} \cdot \frac{\partial \ell_\theta}{\partial \theta}^\top}^{\text{random var.}} \cdot \underbrace{\frac{\partial \theta}{\partial \gamma}^\top}_{p \times p} \right) \\ &= \frac{\partial \theta}{\partial \gamma} J_\theta(\theta) \frac{\partial \theta}{\partial \gamma}^\top \end{aligned}$$

$$\begin{aligned} f_\gamma(\gamma) &= |J_\gamma(\gamma)|^{1/2} = \left| \frac{\partial \theta}{\partial \gamma} \right|^{1/2} |J_\theta(\theta)|^{1/2} \left| \frac{\partial \theta}{\partial \gamma} \right|^{1/2} \\ &= \left| \frac{\partial \theta}{\partial \gamma} \right| f_\theta(\theta) \end{aligned}$$

Ex.: $X \sim \text{bin}(n, p)$, $\ell(p) = x \ln p + (n-x) \ln(1-p)$

$$\frac{d\ell}{dp} = \frac{x}{p} - \frac{n-x}{1-p}$$

$$\frac{d^2\ell}{dp^2} = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}$$

$$J(p) = E\left(-\frac{d^2\ell}{dp^2}\right) = \frac{np}{p^2} + \frac{n(1-p)}{(1-p)^2} = \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)}$$

Jeffreys prior

$$f(p) = (J(p))^{-\frac{1}{2}} \propto p^{-\frac{1}{2}} (1-p)^{-\frac{1}{2}} = p^{\frac{1}{2}-1} (1-p)^{\frac{1}{2}-1}$$

i.e.

$p \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$
also called the arcsine distribution. $F(p) = \frac{\sin^{-1}(2p-1)}{\pi} + \frac{1}{2}$

Sanity check of invariance property:

Reparametrized model, $\theta = \ln \frac{p}{1-p}$

$$\ell(\theta) = x\theta + n \ln(1-p)$$

$$= x\theta + n \ln\left(\frac{1}{1+e^\theta}\right)$$

$$= x\theta - n \ln(1+e^\theta)$$

$$\frac{d\ell}{d\theta} = x - \frac{n e^\theta}{1+e^\theta}$$

$$\frac{d^2\ell}{d\theta^2} = -\frac{n e^\theta}{(1+e^\theta)^2}$$

$$f(\theta) = (J(\theta))^{-\frac{1}{2}} \propto \frac{e^{\theta/2}}{1+e^\theta}$$

Via transformation formula:

$$f(\theta) = f(p) \left| \frac{dp}{d\theta} \right| = p^{-\frac{1}{2}} (1-p)^{-\frac{1}{2}} \frac{e^\theta}{(1+e^\theta)^2} = \frac{e^{\theta/2}}{1+e^\theta}$$

If we instead were to observe $Y \sim \text{geom}(p)$ (where p is the same parameter), the Jeffreys prior (our prior beliefs about p) would have needed to be different!

Criticism: Absurd that prior beliefs is determined (via the likelihood) by the type of data we happen to have obtained.

Summary part 1, sampling methods

Sampling method	When to use	Normalized $f(x)$	Pros	Cons
Inversion	F^{-1} avail F avail	(yes)	Fast	Need to solve $F(x)=u$ numerically
Transformation from simpler RV.	Normal (Box-Muller), location-scale families	(yes)	Fast	Simple transformation may not exist
Rejection sampling	c and $g(x)$ available	No		Slow in high dimensions Finding $g(x)$ similar to $f(x)$ s.t. $f(x)/g(x)$ is bounded
Weighted resampling	c unknown	No		Only approximate
Adaptive rejection sampling	$f(x)$ log-concave	No	Fast	Only in one dimension
Ratio of uniforms	Support of $f(x)$ unbounded, $f(x)$ and $x^2 f(x)$ bounded	No	Fast	
Mixtures	If sampling from $f(x y)$ and $f(y)$ is easy			
Multivariate normal				
Alias method	pmfs with finite support			
Simulating generating process	Sanity checking process			Slower

Estimating expectations, $E(h(x))$

Integration method	When to use	Normalized $f(x)$	Pros	Cons
Analytic	Whenever possible	Yes	Exact	
Numerical Adaptive Gauss Hermite quadrature	Integrand approx. Gaussian	Yes	Fast and often very accurate	Difficult in for high-dimensions
Monte Carlo		No	Unbiased	High variance
Importance sampling		Yes	Unbiased, smaller variance	Finding $g(x)$ similar to $h(x)f(x)$
		No	Consistent	Biased
Antithetic sampling				