Submission:

- A written report in PDF format (*No* MS Word-files). The pair Errol Norqvist and Lillemor Dahlstedt name their report file `proj1-dlne.pdf`.
- An email containing the report file as well as *all* your R-files with a file `proj1.R` that runs your analysis. This email has to be sent to `jonwal@chalmers.se` before Monday 21 Sep, 23:59
- Late submissions do not qualify for bonus points

Instructions on report writing

- Explain carefully all introduced notation: $X = ?$.
- Describe/explain the model.
- The text should be readable without access to the R code; write plain text instead of including R code in the report.
- Include your solutions in the text; do not write "calculations of ? can be found in the R code", or similar.
- When referring to the lecture notes or the book, be specific (i.e. refer to Chapter/which lecture).
- Refer to your figures in the text. Explain colors etc. in the figure captions (a figure caption is almost never to long).
- Motivation and reasoning when it concerns choice of instrumental distributions etc.
- Page limit: 12 pages, 12 points

Computer intensive statistical methods

Lecture 3
Rejection and Importance sampling

September 15, 2015

Jonas Wallin
jonwal@chalmers.se

Chalmers, Gothenburg university

A law of large numbers

Theorem (law of large numbers)

Let $X_1, X_2, \ldots, X_N$ be independent identically distributed random variables and h a function such that $\mathbb{V}[h(X_i)] < \infty$ for all i. If $N \to \infty$ then

$$\tau_N \overset{def.}{=} \frac{1}{N} \sum_{i=1}^{N} h(X_i) \xrightarrow{\mathbb{P}} \mathbb{E}(h(X)).$$

Convergence in probability

$$\frac{1}{N}\sum_{i=1}^{N} h(X^{(i)}) \xrightarrow{\mathbb{P}} \mathbb{E}_f(h(X)),$$

implies that: for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} h(X^{(i)}) - \int_{\mathsf{X}} h(x)f(x)\,\mathrm{d}x\right| \geq \epsilon\right) \to 0 \quad \text{as} \quad N \to \infty.$$

Proof of LLN (helping Lemma)

Lemma (Chebyshev's inequality)

Let $Z \geq 0$ be a r.v. with $\mathbb{E}Z^2 < \infty$ then

$$\mathbb{P}(Z \geq \epsilon) \leq \frac{\mathbb{E}[Z^2]}{\epsilon^2}.$$

Proof.

$$\begin{aligned}\mathbb{P}(Z \geq \epsilon) &= \int_\epsilon^\infty f_Z(z)\mathrm{d}z \leq \int_\epsilon^\infty \left(\frac{z}{\epsilon}\right)^2 f_Z(z)\mathrm{d}z \\ &\leq \int_0^\infty \left(\frac{z}{\epsilon}\right)^2 f_Z(z)\mathrm{d}z = \frac{\mathbb{E}[Z^2]}{\epsilon^2}\end{aligned}$$

□

Proof of LLN

Proof.

First:

$$\mathbb{E}[\tau_N] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} h(X^{(i)})] = \frac{1}{N}\mathbb{E}[\sum_{i=1}^{N} h(X)] = \mathbb{E}_f[h(X)].$$

Second we bound the variance: Denote $S_N = \sum_{i=1}^{N} h(X_i), (\tau_N = \frac{S_N}{N})$ and $\tau = \mathbb{E}[h(X)]$, then

$$\mathbb{V}[\tau_N] = \mathbb{E}[\frac{1}{N^2}(S_N - N\tau)^2] =$$

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\mathbb{E}[(h(X_i) - \tau)(h(X_k) - \tau)].$$

Proof of cont

Proof.

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\mathbb{E}[(h(X_i)-\tau)(h(X_k)-\tau)] =$$

$$\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}[(h(X_i)-\tau)^2] = \frac{1}{N}\mathbb{V}[h(X)].$$

Then use Chebyshev □

Confidence bounds

Last time we noted that the central limit theorem (CLT) implies

$$\sqrt{N}\left(\tau_N - \tau\right) \xrightarrow{\text{d.}} \mathcal{N}(0, \sigma^2(h)), \quad \text{as } N \to \infty,$$

where

$$\sigma^2(h) = \mathbb{V}\left(h(X)\right).$$

Consequently, the two-sided confidence interval

$$\mathcal{I}_\alpha = \left(\tau_N - \lambda_{\alpha/2}\frac{\sigma(h)}{\sqrt{N}}, \tau_N + \lambda_{\alpha/2}\frac{\sigma(h)}{\sqrt{N}}\right),$$

where λ_p denotes the p-quantile of the standard normal distribution, covers τ with (approximate) probability $1 - \alpha$.

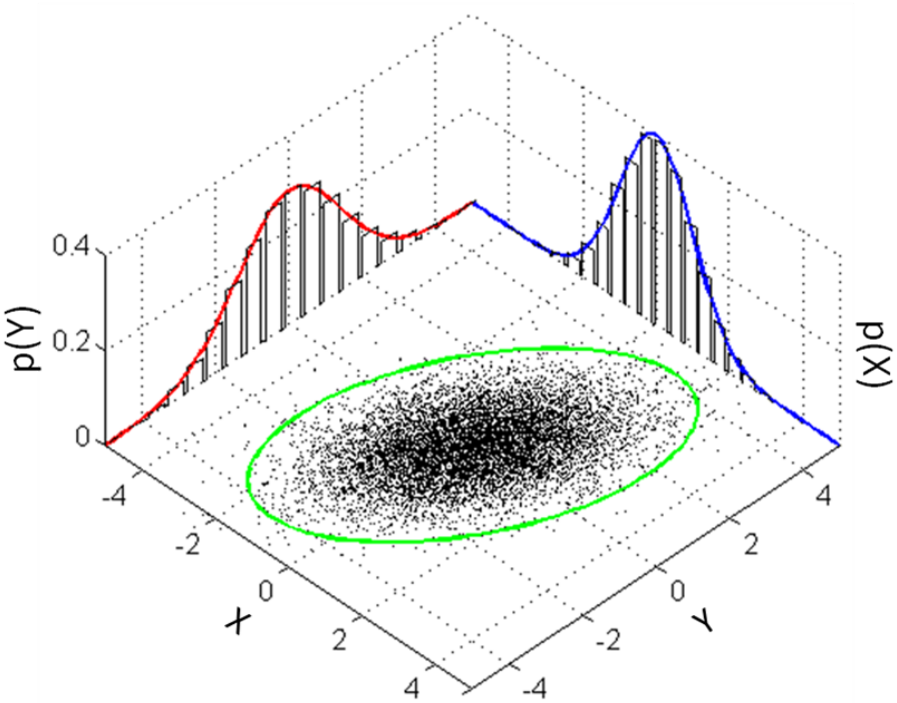
$\sigma^2(h)$ can be approximated with

$$\sigma_N^2(h) \overset{\text{def.}}{=} \frac{1}{N-1}\sum_{i=1}^N \left(h(X_i) - \frac{1}{N}\sum_{l=1}^N h(X_l)\right)^2.$$

The Multivariate Normal

The Multivariate Normal distribution denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has density ($x \in \mathbb{R}^n$)

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{-n/2}} |\Sigma|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$



The Multivariate delta method

For a given estimand $\boldsymbol{\tau} \in \mathbb{R}^d$, one is often interested in estimating $\mathbf{g}(\boldsymbol{\tau})$ for some further function $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^m$. One can estimate $\mathbf{g}(\boldsymbol{\tau})$ using $\mathbf{g}(\boldsymbol{\tau}_N)$

Under suitable assumptions on $\mathbf{g}$ and if

$$\sqrt{N}\left(\boldsymbol{\tau}_{\mathbf{N}} - \boldsymbol{\tau}\right) \xrightarrow{\text{d.}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(h))$$

then

$$\sqrt{N}\left(\mathbf{g}(\boldsymbol{\tau}) - \mathbf{g}(\boldsymbol{\tau}_{\mathbf{N}})\right) \xrightarrow{\text{d.}} \mathcal{N}(\mathbf{0}, \nabla\mathbf{g}(\boldsymbol{\tau})^T \boldsymbol{\Sigma}(h) \nabla\mathbf{g}(\boldsymbol{\tau})).$$

The delta method, an example

For $(X_i)_{i=1}^N$ iid drawn from f then by the multivariate delta method:

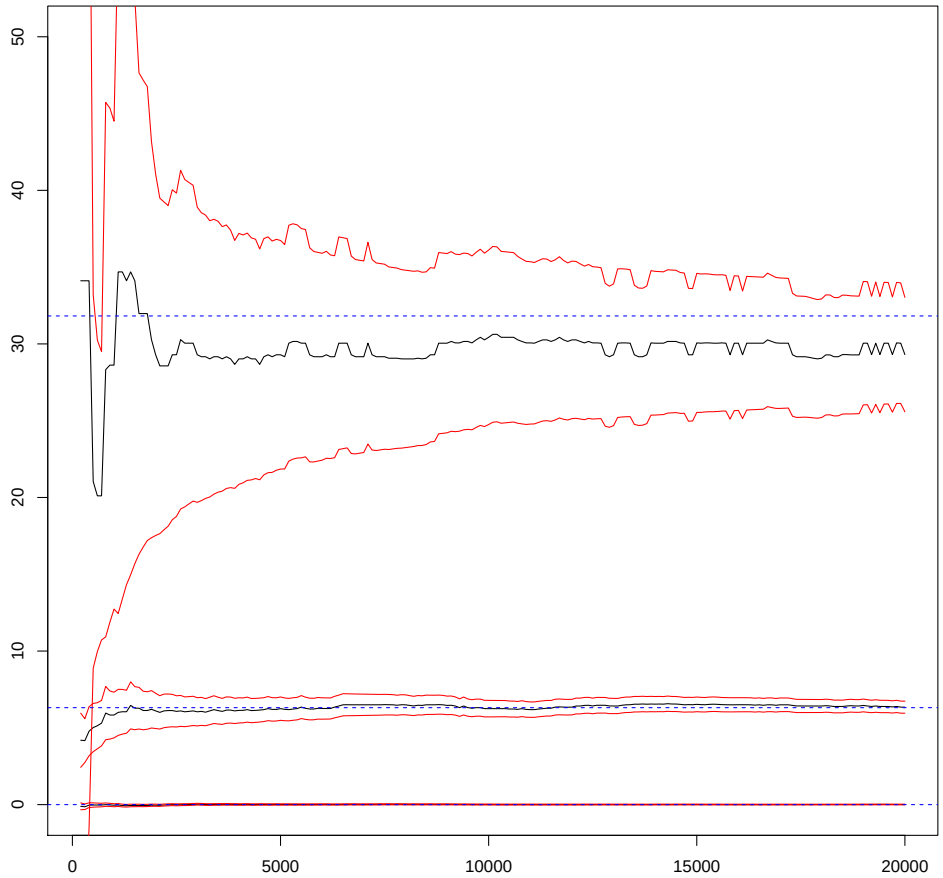
$$\sqrt{N}\left(\begin{bmatrix} X_{([Np_1])} \\ X_{([Np_2])} \\ \vdots \\ X_{([Np_m])} \end{bmatrix} - \begin{bmatrix} x_{p_1} \\ x_{p_2} \\ \vdots \\ x_{p_m} \end{bmatrix}\right) \xrightarrow{\text{d.}} \mathcal{N}\left(\mathbf{0}, \nabla\mathbf{g} \begin{bmatrix} p_1(1-p_1) & \dots & p_1(1-p_m) \\ \vdots & & \vdots \\ p_1(1-p_m) & \dots & p_m(1-p_m)) \end{bmatrix} \nabla\mathbf{g}\right),$$

where $x_p = F^{-1}(p)$ (that is the p-quantile) and $\nabla\mathbf{g}$ is a diagonal matrix with $(\nabla\mathbf{g})_{ii} = \frac{1}{f(x_{p_i})}$.

One can note that the correlation for the quantiles is independent of f.

Cauchy quantiles

The 50%, 95%, 99% quantile with confidence bounds from a Cauchy distribution.



Cauchy quantiles (cont.)

From the delta method we get the covariance matrix

$$\nabla \mathbf{g} \begin{bmatrix} p_1(1-p_1) & \dots & p_1(1-p_m) \\ \vdots & & \vdots \\ p_1(1-p_m) & \dots & p_m(1-p_m)) \end{bmatrix} \nabla \mathbf{g} \approx \begin{bmatrix} 4 & 10 & 500 \\ 10 & 782 & 3883 \\ 500 & 3883 & 100374 \end{bmatrix}$$

Indicating that pure Monte Carlo is a bad for rare events.

The inversion method

Let $\mathcal{U}[0,1]$ pseudo-random numbers U and want to generate random numbers X from a univariate distribution with distribution function F.

Define the general inverse $F^{\leftarrow}(u) \stackrel{\text{def.}}{=} \inf\{x \in \mathbb{R} : F(x) \geq u\}$ and

draw $U \sim \mathcal{U}[0,1]$
set $X \leftarrow F^{\leftarrow}(U)$
return X

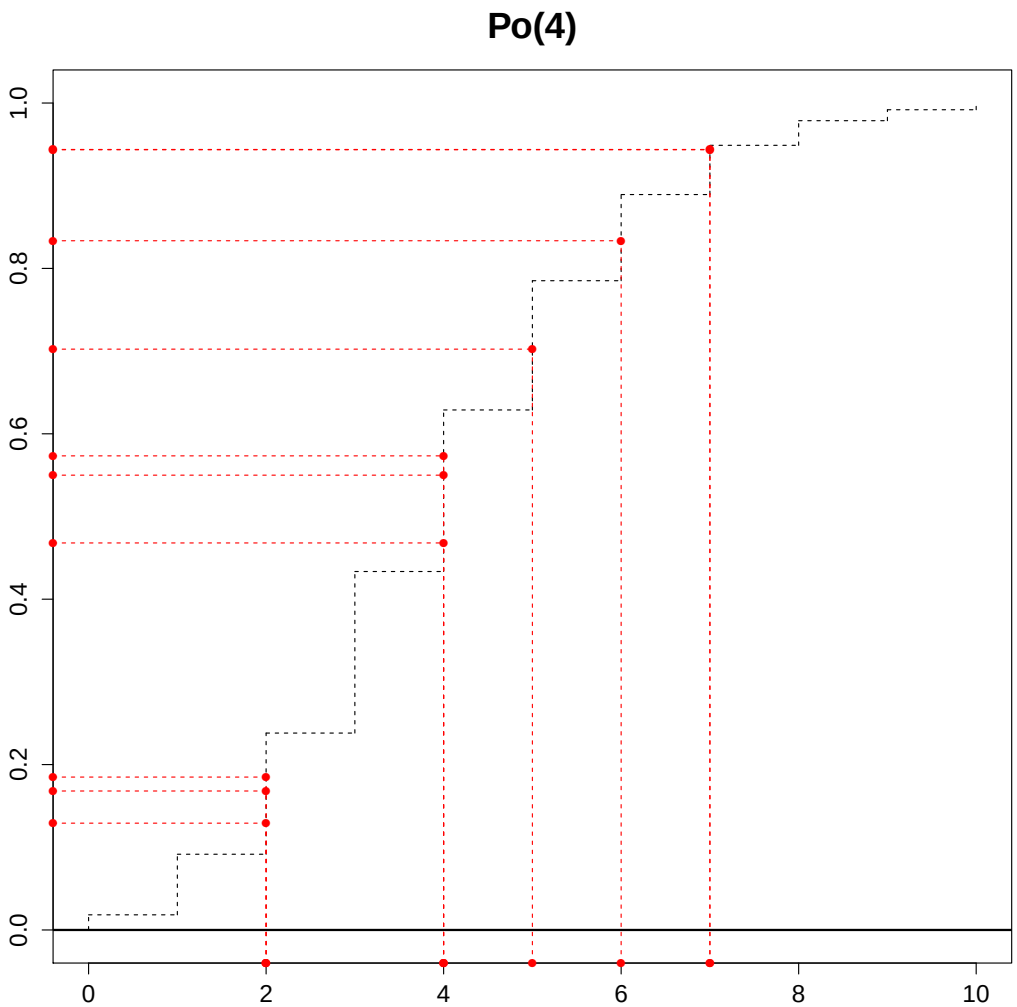
One may now prove the following.

Theorem (Inverse method)

The output X has distribution function F.

The inversion method

$$F^{\leftarrow}(u) \stackrel{\text{def.}}{=} \inf\{x \in \mathbb{R} : F(x) \geq u\}$$



univariate Transformation method

The continuous inversion method is a special case of a broader class of methods based on the transformation of random numbers. From the change of variable

Theorem (Transfromation method)

Let X be a random variable with density f_X and support $\mathbb{X} \subset \mathbb{R}$. Let g be a differentiable 1-1 function from $\mathbb{X}$ to $\mathbb{Y} \subset \mathbb{R}$ with inverse g^{-1}. Then $Y = g(X)$ is a random variable with support $\mathbb{Y}$ and density given by

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

Transformation method

Theorem (Transfromation method)

Let $\mathbf{X}$ *be a random variable with density* f_X *and support* $\mathbb{X} \subset \mathbb{R}^d$. *Let* g *be a differentiable 1-1 function from* $\mathbb{X}$ *to* $\mathbb{Y} \subset \mathbb{R}^d$ *with inverse* g^{-1}. *Then* $\mathbf{Y} = g(\mathbf{X})$ *is a random variable with support* $\mathbb{Y}$ *and density given by*

$$f_Y(\mathbf{y}) = f_X(g^{-1}(\mathbf{y}))\,|J(\mathbf{y})|$$

where $J(\mathbf{y})$ *is the Jacobian of the transformation.*

Recall That is the determinant of the matrix with (i, j) element

$$J(\mathbf{y})_{ji} = \frac{d}{d\mathbf{y}_j} g_i^{-1}(\mathbf{y})$$

where $g_i^{-1}(\mathbf{y})$ is the function so $x_i = g_i^{-1}(\mathbf{y})$.

sampling Multivariate Normal distribution

Let $\mathbf{X}$ be a vector of d independent $N(0,1)$ variables and $\mathbf{R}$ the (upper) Cholesky factorization of $\mathbf{\Sigma}$. Then

$$\mathbf{Y} = \mu + \mathbf{R}^T\mathbf{X},$$

then $\mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbf{\Sigma})$.

```
x <- rnorm(d)
R <- chol(Sigma)
y <- mu + t(R)%*%x
```

The agenda of the day

- Sampling from non standard distribution (rejection method)
- effective Monte Carlo integration (importance sampling)
- adaptive Importance sampling

Rejection sampling

The inversion method works well when it can be applied (although often not best), but what do we do if, e.g., $f(x) \propto \exp(\cos^2(x))$, $x \in (-\pi/2, \pi/2)$? Here we cannot find an inverse .

In the continuous case the following (somewhat magic!) algorithm saves the day. Let f and g be densities on $\mathbb{R}^d$ for which there exists a constant $M < \infty$ such that $f(x) \leq Mg(x)$ for all $x \in \mathbb{R}^d$; then

repeat
  draw $X^* \sim g$
  draw $U \sim \mathcal{U}[0, 1]$
until $U \leq \frac{f(X^*)}{Mg(X^*)}$
$X \leftarrow X^*$
return X

Example

We wish to simulate $f(x) = \exp(\cos^2(x))/c$, $x \in (-\pi/2, \pi/2)$, where $c = \int_{-\pi/2}^{\pi/2} \exp(\cos^2(z))\, \mathrm{d}z = \pi e^{1/2} I_o(1/2)$ is the normalizing constant.

However, since for all $x \in (-\pi/2, \pi/2)$,

$$f(x) = \frac{\exp(\cos^2(x))}{c} \leq \frac{e}{c} = \underbrace{\frac{e\pi}{c}}_{M} \times \underbrace{\frac{1}{\pi}}_{g}$$

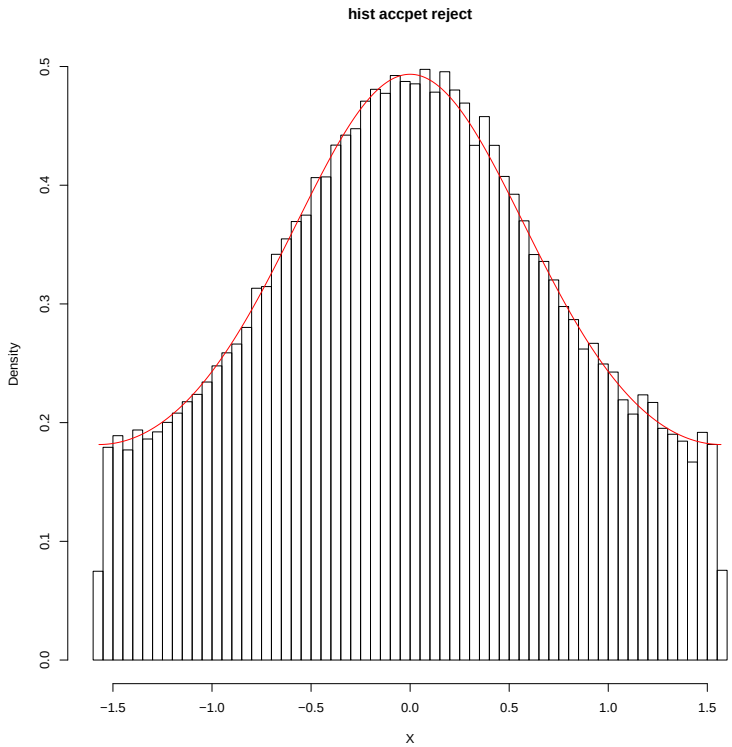
where g is the density of $\mathcal{U}[-\pi/2, \pi/2]$.

we may use rejection sampling where a candidate $X^* \sim \mathcal{U}[-\pi/2, \pi/2]$ is accepted if

$$U \leq \frac{f(X^*)}{Mg(X^*)} = \frac{\exp(\cos^2(X^*))/c}{e/c} = \exp(\cos^2(X^*) - 1).$$

Example

```
ratio <- function(x){f(x)/exp(1)}
for(i in 1:N)
{
  X_star <- -pi/2 + pi*runif(1)
  while( runif(1) > ratio(X_star)){X_star <- -pi/2 + pi*runif(1)}
  X[i] <- X_star
}
```



Rejection sampling (cont.)

Theorem (Rejection sampling*)

The output X of the rejection sampling algorithm has density function f.

Moreover:

Theorem

The expected number of trials needed before acceptance is M.

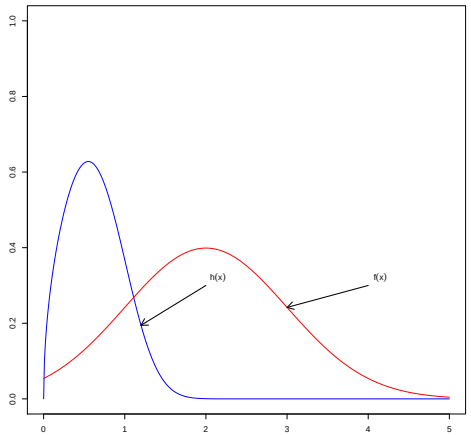
Consequently, M should be chosen as small as possible.

Further, we only need to know f,g up to a normalizing constant!

Problems with MC integration

We may run into problem with MC integration if:

- it is hard to sample from f or
- if the integrand h and the density f are dissimilar; in this case we will end of with a lot of draws where the integrand is small, and consequently only a few draws will contribute to the estimate. This gives a large variance.



These problems can often be solved using importance sampling.

Importance sampling (IS)

The basis of importance sampling is to take an instrumental density g on χ such that $g(x) = 0 \Rightarrow f(x) = 0$ and rewrite the integral as

$$\tau = \mathbb{E}_f\left(h(X)\right) = \int_\chi h(x) f(x)\,\mathrm{d}x = \int_{f(x)>0} h(x) f(x)\,\mathrm{d}x$$

$$= \int_{g(x)>0} h(x)\frac{f(x)}{g(x)} g(x)\,\mathrm{d}x = \mathbb{E}_g\left(h(X)\frac{f(X)}{g(X)}\right) = \mathbb{E}_g\left(h(X)\omega(X)\right),$$

where

$$\omega : \{x \in \chi : g(x) > 0\} \ni x \mapsto \frac{f(x)}{g(x)}$$

is the so-called importance weight function.

Importance sampling (cont.)

We may now estimate $\tau = \mathbb{E}_g(h(X)\omega(X))$ using standard MC:

for $i = 1 \to N$ **do**
  draw $X_i \sim g$
end for
set $\tau_N \leftarrow \sum_{i=1}^{N} h(X_i)\omega(X_i)/N$
return τ_N

Here,

$$\mathbb{V}(\tau_N) = \frac{1}{N}\mathbb{V}_g(h(X)\omega(X)).$$

Importance sampling variance

Choosing what g to use, we are looking for the g such that

$$g^* = \arg\min_g \mathbb{V}_g(h(X)\omega(X)).$$

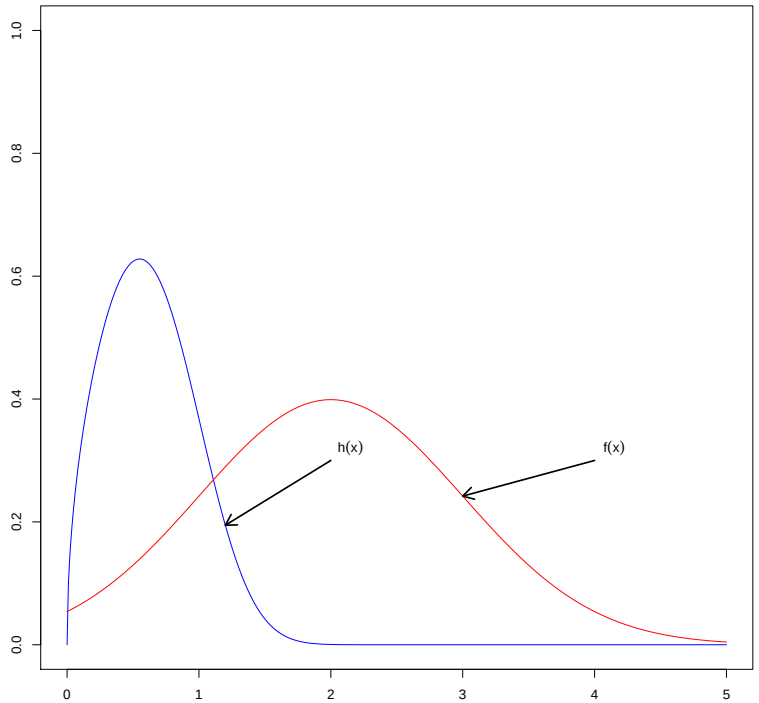
Which can be written as

$$g^*(x) \propto |h(x)|f(x).$$

Example: A normal expectation

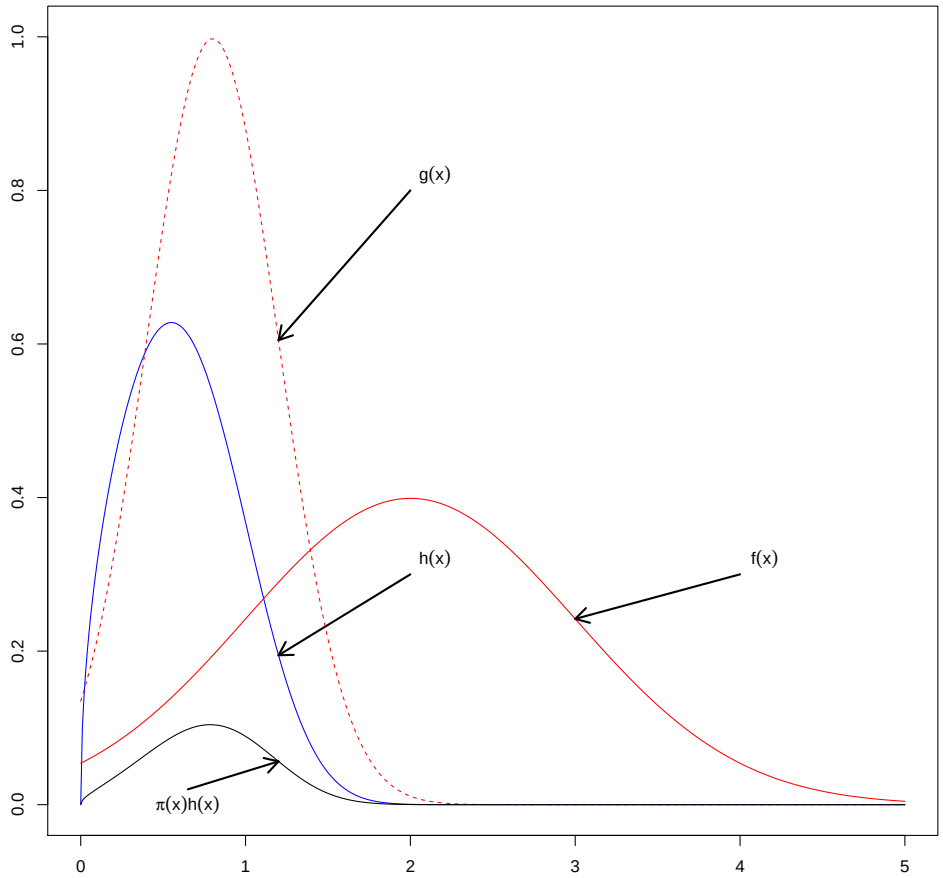
Let X have $\mathcal{N}(2,1)$ distribution and try to compute

$$\tau = \mathbb{E}\left(\mathbb{I}(X \geq 0)\sqrt{X}\exp(-X^3)\right) = \int \underbrace{\mathbb{I}(x \geq 0)\sqrt{x}\exp(-x^3)}_{=h(x)} \underbrace{\mathcal{N}(x;2,1)}_{=f(x)} \, \mathrm{d}x,$$



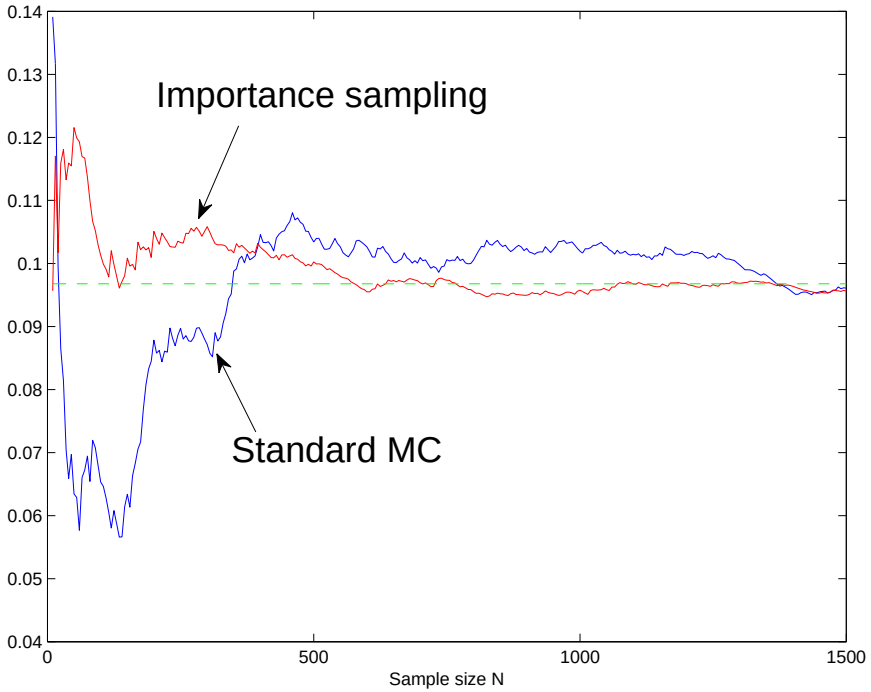
Example: A normal expectation (cont.)

Thus, standard MC will lead to a waste of computational power. Better is to use IS with g being a scale-location-transformed normal-distribution:



Example: A normal expectation (cont.)

```
mu    <- 0.8
sigma <- 0.4
omega <- function(x,mu,sigma){ dnorm(x,2,1)/dnorm(x, mu, sigma)}
X <- sigma * rnorm(N) + mu
tau_IS <- cumsum(h(X) * omega(X, mu, sigma))
```



Importance sampler proposal distribution

- Rejection sampling throws away a lot of samples
- Performance dependence on K which must be choosen.
- Importance sampling works for all distributions given they have the same support?

What happens if the importance sampler proposal has lighter tails then the function that we are interested in?

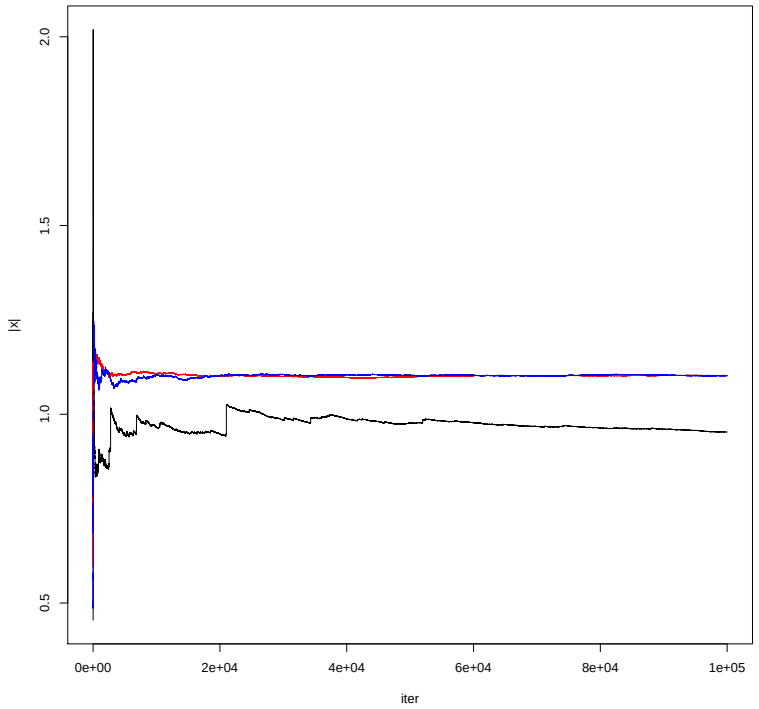
$$\mathbb{E}_g[h(X)^2\omega(X)^2] = \int h(x)^2\omega(x)^2 g(x)\,\mathrm{d}x = \int h(x)^2 \frac{f(x)^2}{g(x)}\,\mathrm{d}x,$$

If the tails of g is lighter then π then the variance could be ∞.

Importance sampler proposal distribution, example

An comparison of different proposals suppose that $f(x) = |x|$ and $f(x) = f(x; df = 3)$. We compare three different proposals:

$$g_1(x) = \mathcal{N}(x; 0, 1),$$
$$g_2(x) = Cauchy(x; 0, 1),$$
$$g_3(x) = t(x; df = 3) \quad \text{(regular Monte Carlo)}$$



Adaptive importance sampling

Suppose we have choose a instrumental distribution how do we set the parameters?

- Ideally we find

$$\arg\min_{\theta} \mathbb{V}_{g_\theta} \left[h(X)\omega(X;\theta) \right].$$

  However, this is difficult, that is seldom an explicit solution, and the numerical solution is often an unstable estimation.

- Instead, we try to find a distribution close to the optimal choice g^*.

Kullback-Leibler diverence

So we need to define what close is for two densities. A popular "distance" (not actualy a distance) is the Kullback-Leibler divergence.

- $$\mathcal{D}(f,g) = \mathbb{E}_f \ln \frac{f(X)}{g(X)}.$$

  If $\mathcal{D}(f,g)$ is small if f and g are considered close.
- Further $\mathcal{D}(f,g) = 0$ iff $f(x) = g(x)$. and $\mathcal{D}(f,g) = \infty$ if $g(x) = 0$ and $f(x) > 0$.

Kullback-Leibler diverence, Likelihood

A connection to likelihood estimation is:

- Assume g_θ is parametric distribution. And we want to find the closest distribution in KL sense.
- Approximate the integral $\mathcal{D}(f, g)$ with MC version: $\mathcal{D}_N(f, g_\theta) = \frac{1}{n}\sum_{i=1}^n \log(\frac{f(X_i)}{g_\theta(X_i)})$, where X_i iid samples from f.
- To find the closet distribution solve

$$\arg\min_\theta \mathcal{D}_N(f, g_\theta) = \arg\min_\theta -\sum_{i=1}^n \log(g_\theta(X_i)) = \arg\max_\theta \prod_{i=1}^n g_\theta(X_i).$$

This is equivalent to the Maximum likelihood estimator and independent of f.

KL for importance sampling

Recall that the ideal density is $g^*(x) = |h(x)|f(x)$.

- $$\mathcal{D}(g^*, g_{\boldsymbol{\Theta}}),$$

  is equivalent to maximizing

  $$\mathbb{E}_f |h(X)| \ln g_{\boldsymbol{\Theta}}(X).$$

  where the maximum can typically be found using

  $$\mathbb{E}_\pi |h(X)| \nabla_\Theta \ln g_{\boldsymbol{\Theta}}(X) = 0.$$

KL for importance sampling cont.

Of course we cant analytically find

$$\mathcal{D}(g^*, g_{\boldsymbol{\Theta}}) = \mathbb{E}_f |h(X)| \nabla_{\Theta} \ln g_{\boldsymbol{\Theta}}(X) = 0.$$

However often we can approximate the integral in the IS, which often gives a closed form solution

$\tau_N = 0$
for $i = 1 \to N$ **do**
  draw $X_i \sim g_{\Theta^{(i-1)}}$
  $\tau_N \leftarrow \tau_N + h(X_i)\omega(X_i; \Theta^{(i-1)})$
  $\Theta^{(i)} \leftarrow \arg_\theta \sum_{j=1}^{i} h(X_j)\omega(X_j; \Theta^{(j-1)}) \nabla_\theta \ln g_\Theta(X_j) = 0$
end for
return τ_N / N

Kullback-Leibler diverence example

Let X have $\mathcal{N}(2,1)$ distribution and try to compute

$$\tau = \mathbb{E}\left(\mathbb{I}(X \geq 0)\sqrt{X}\exp(-X^3)\right) = \int \underbrace{\mathbb{I}(x \geq 0)\sqrt{x}\exp(-x^3)}_{=h(x)} \underbrace{\mathcal{N}(x;2,1)}_{=f(x)} \, \mathrm{d}x,$$

with instrumental distribution $g(x) = \mathcal{N}(x;\mu,\sigma)$. Thus we need to find $\boldsymbol{\Theta} = \{\mu, \sigma\}$.