MATH 559: BAYESIAN THEORY AND METHODS
IMPORTANCE SAMPLING AND VARIANCE REDUCTION

Importance Sampling: We may write

$$I(g) \equiv \mathbb{E}_f[g(X)] = \int g(x)f(x)dx = \int \frac{g(x)f(x)}{f_0(x)} f_0(x)dx = \mathbb{E}_{f_0}\left[\frac{g(X)f(X)}{f_0(X)}\right]$$

provided that the support of $f_0(x)$ includes the support of $f(x)$. This suggests the *Importance Sampling* (IS) strategy where $X_1, \ldots, X_N$ are sampled independently from f_0, with the estimator

$$\widehat{I}_N^{(f_0)}(g) = \frac{1}{N}\sum_{i=1}^{N} \frac{g(X_i)f(X_i)}{f_0(X_i)}.$$

The choice $f_0(x) \equiv f(x)$ for all x returns the original MC estimator, $\widehat{I}_N(g)$. We might use IS if f is hard to sample from whereas f_0 is straightforward, or if the variance of $\widehat{I}_N^{(f_0)}(g)$ is smaller than that of $\widehat{I}_N(g)$. We have that

$$\text{Var}_{f_0}\left[\widehat{I}_N^{(f_0)}(g)\right] = \frac{1}{N}\text{Var}_{f_0}\left[\frac{g(X)f(X)}{f_0(X)}\right] = \frac{1}{N}\int \left(\frac{g(x)f(x)}{f_0(x)} - I(g)\right)^2 f_0(x)dx = \frac{1}{N}\int \left\{\frac{g(x)f(x)}{f_0(x)}\right\}^2 f_0(x)dx - \frac{\{I(g)\}^2}{N}$$

which shows that the variance is finite if and only if

$$\int \left\{\frac{g(x)f(x)}{f_0(x)}\right\}^2 f_0(x)\,dx < \infty.$$

We have that by Jensen's inequality

$$\int \left\{\frac{g(x)f(x)}{f_0(x)}\right\}^2 f_0(x)\,dx \ge \left\{\int \frac{|g(x)|f(x)}{f_0(x)} f_0(x)\,dx\right\}^2 = \left\{\int |g(x)|f(x)\,dx\right\}^2$$

and so we have that the right-hand side is a lower bound on the left-hand side. We can make the two sides equal, and therefore achieve the lower bound by setting

$$f_0(x) = \frac{|g(x)|f(x)}{\int |g(t)|f(t)\,dt}$$

However, this is infeasible as we cannot explicitly compute the pdf in most cases. It does inform us that we should choose $f_0(x)$ to be close (in 'shape') to $|g(x)|f(x)$.

Note also that the statistic

$$\frac{1}{N}\sum_{i=1}^{N}\frac{f(X_i)}{f_0(X_i)} \xrightarrow{a.s.} \int \frac{f(x)}{f_0(x)} f_0(x)\,dx = 1$$

so we may instead use the estimator

$$\frac{\displaystyle\sum_{i=1}^{N}\frac{g(X_i)f(X_i)}{f_0(X_i)}}{\displaystyle\sum_{j=1}^{N}\frac{f(X_j)}{f_0(X_j)}} = \sum_{i=1}^{N} w(X_i)g(X_i) \qquad w(X_i) = \frac{\dfrac{f(X_i)}{f_0(X_i)}}{\displaystyle\sum_{j=1}^{N}\frac{f(X_j)}{f_0(X_j)}} \qquad i = 1, \ldots, N.$$

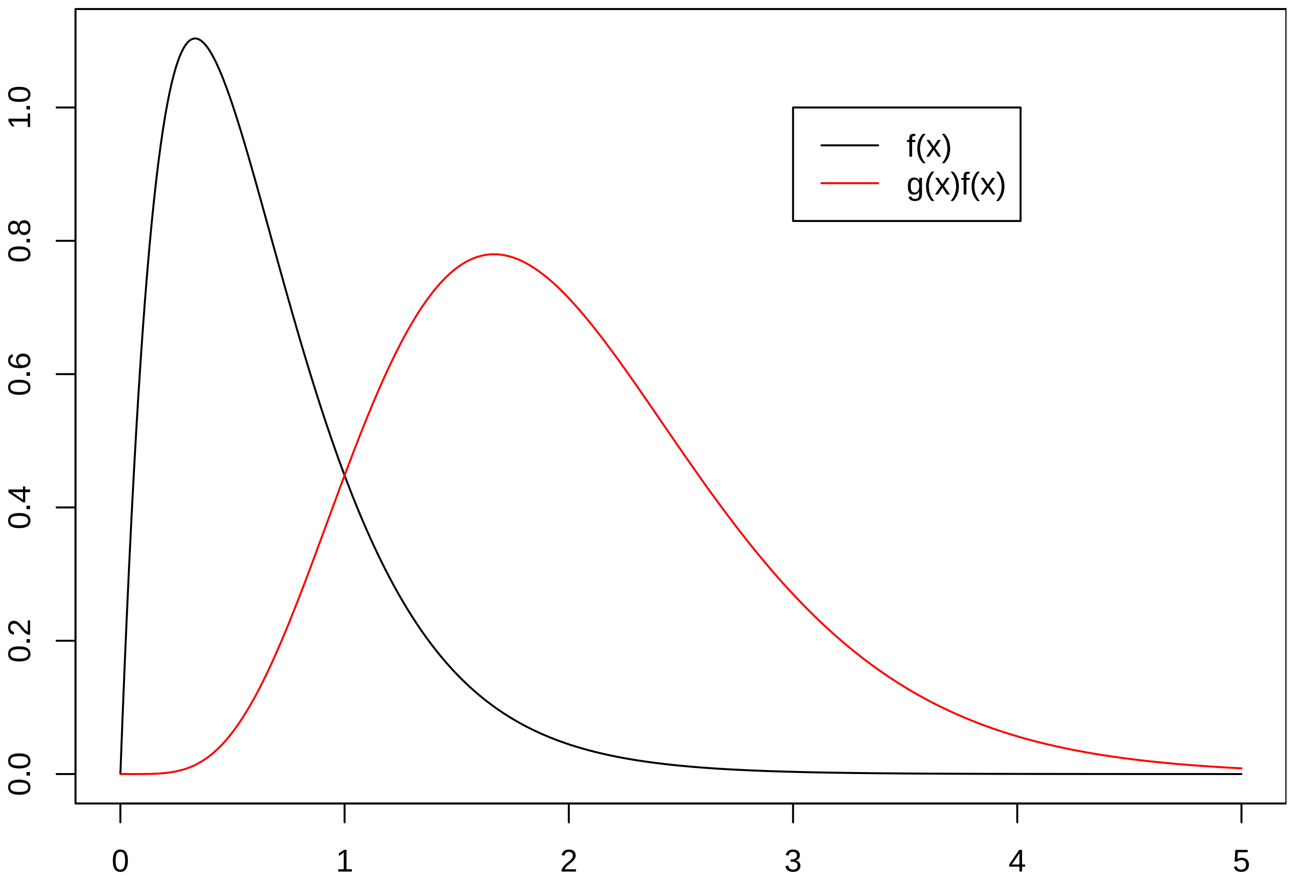
This approach can be useful if the densities $f(x)$ and $f_0(x)$ are only known up to proportionality.

EXAMPLE: Suppose $X \sim Gamma(2,3)$, and aim to compute $\mathbb{E}[X^4]$. For $r > 0$, with $X \sim Gamma(\alpha, \beta)$

$$\mathbb{E}[X^r] = \frac{1}{\beta^r}\frac{\Gamma(\alpha + r)}{\Gamma(\alpha)}$$

so therefore $\mathbb{E}[X^4] = (5 \times 4 \times 3 \times 2)/3^4 = 1.481481$.

```
al<-2;be<-3;r<-4
true.val<-(gamma(al+r)/gamma(al))/be^r
xv<-seq(0,5,by=0.01)
yv<-dgamma(xv,al,be)
yv4<-xv^4
par(mar=c(3,3,1,0))
plot(xv,yv,type='l',xlab='y',ylab=' ')
lines(xv,yv*yv4,col='red')
legend(3,1,c('f(x)','g(x)f(x)'),lty=1,col=c('black','red'))
```



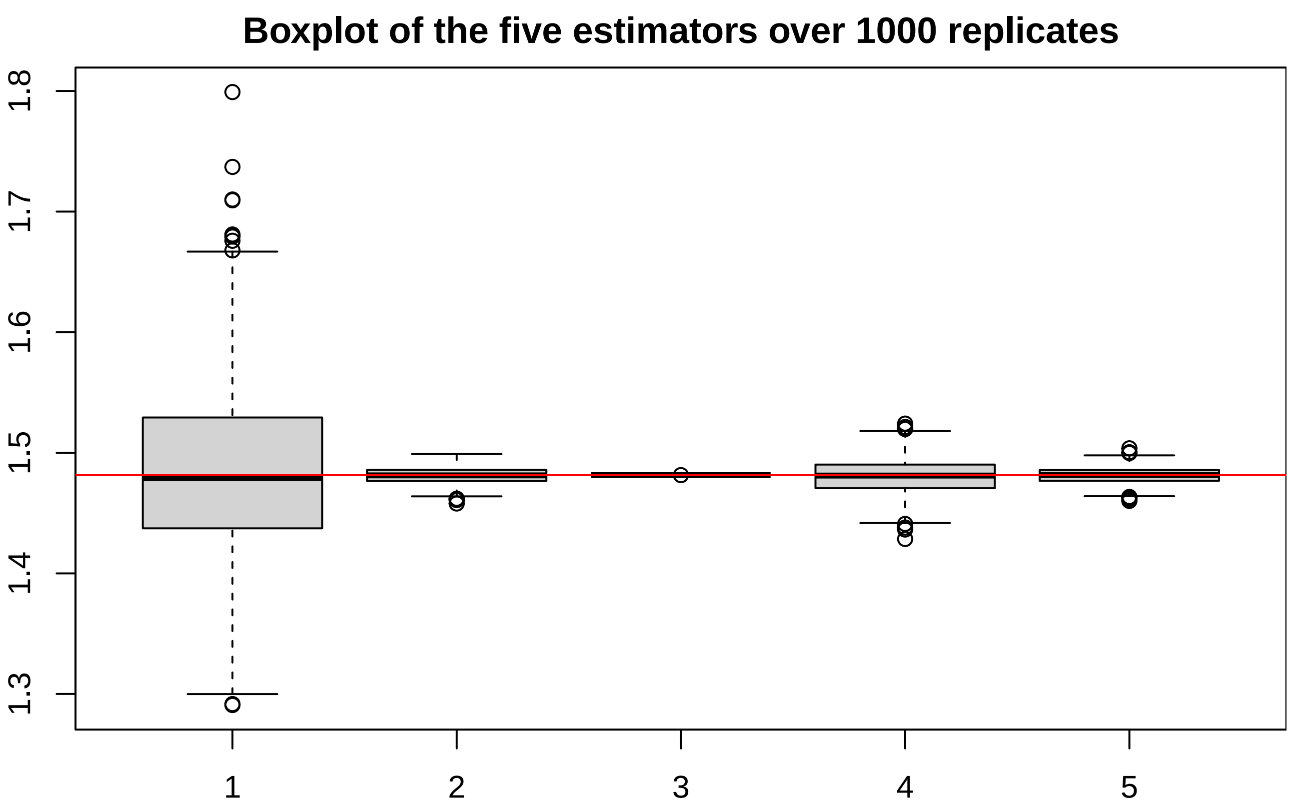
We test the importance sampling estimator with the following choices of f_0

- $f_0(x) \equiv Gamma(2,3)$ (that is, ordinary MC estimation);
- $f_0(x) \equiv Gamma(5,2)$;
- $f_0(x) \equiv Gamma(6,3)$;
- $f_0(x) \equiv Normal(2,2)$;
- $f_0(x) \equiv Student(5)$ centered at $y = 2$;

Replicate runs show the variability in each case

```
set.seed(64)
N<-10000
nreps<-1000
IS.ests<-matrix(0,nrow=nreps,ncol=5)
for(irep in 1:nreps){
    X1<-rgamma(N,2,3)
    IS.ests[irep,1]<-mean(X1^4)
    X2<-rgamma(N,5,2)
    IS.ests[irep,2]<-mean(X2^4*dgamma(X2,2,3)/dgamma(X2,5,2))
```

```
    X3<-rgamma(N,6,3)
    IS.ests[irep,3]<-mean(X3^4*dgamma(X3,2,3)/dgamma(X3,6,3))
    X4<-rnorm(N,2,2)
    IS.ests[irep,4]<-mean(X4^4*dgamma(X4,2,3)/dnorm(X4,2,2))
    X5<-rt(N,df=5)+2
    IS.ests[irep,5]<-mean(X5^4*dgamma(X5,2,3)/dt(X5-2,df=5))
}
par(mar=c(3,3,2,0))
boxplot(IS.ests);title('Boxplot of the five estimators over 1000 replicates')
abline(h=true.val,col='red')
```




```
#True.value
true.val

+ [1] 1.481481

apply(IS.ests,2,mean)   #Mean of estimator

+ [1] 1.484467 1.481253 1.481481 1.480800 1.481221

N*apply(IS.ests,2,var)  #Variance of estimator

+ [1] 50.954405117578083661555865546688437461 9
+ [2]  0.463469432197713027132834895382984541 4
+ [3]  0.000000000000000000000000000000493531 6
+ [4]  2.196217990126137831197183913900516927 2
+ [5]  0.427657376311393888634881932375719770 8
```

Here it is evident that the IS estimators perform much better (in terms of estimator variance) than the MC estimator. Note that for the case $f_0(x) \equiv Gamma(6,3)$ we have exactly matched the integrand

$$f_0(x) \propto x^{6-1}\exp\{-3x\} = x^4 x^{2-1}\exp\{-3x\} \propto g(x)f(x).$$

Rejection Sampling: Rejection sampling is a form of direct sampling from $f(x)$ that uses an alternate distribution f_0 for sampling which can be used under the condition

$$\frac{f(x)}{f_0(x)} < M < \infty$$

for all y. To apply rejection sampling

(i) generate $X \sim f_0(x)$ and $U \sim Uniform(0,1)$ independently

(ii) **accept** X as a sample from $f(x)$ if

$$U \leq \frac{f(X)}{M f_0(X)}$$

and **reject** X and return to (i) if

$$U > \frac{f(X)}{M f_0(X)}$$

We have that

$$\begin{aligned}\Pr[X \text{ is accepted}] = \Pr\left[U \leq \frac{f(X)}{M f_0(X)}\right] = \int_{-\infty}^{\infty} \left\{ \int_0^{f(x)/(M f_0(x))} du \right\} f_0(x)\, dx &= \int_{-\infty}^{\infty} \frac{f(x)}{M f_0(x)} f_0(x)\, dx \\ &= \frac{1}{M} \int_{-\infty}^{\infty} f(x)\, dx = \frac{1}{M}\end{aligned}$$

and then that for $x \in \mathbb{R}$,

$$\begin{aligned}\Pr[X \leq x | X \text{ is accepted}] = \frac{\Pr[X \leq x, X \text{ is accepted}]}{\Pr[X \text{ is accepted}]} &= M \int_{-\infty}^{x} \left\{ \int_0^{f(t)/(M f_0(t))} du \right\} f_0(t)\, dt \\ &= \int_{-\infty}^{x} \frac{f(t)}{f_0(t)} f_0(t)\, dt = \int_{-\infty}^{x} f(t)\, dt\end{aligned}$$

and therefore the accepted points have distribution $f(x)$.

EXAMPLE: To illustrate the use of rejection sampling, consider the following model:

$$f(x) = \frac{1}{4}\frac{1}{\sigma_1}\phi\left(\frac{y-\mu_1}{\sigma_1}\right) + \frac{3}{4}\frac{1}{\sigma_2}\phi\left(\frac{y-\mu_2}{\sigma_2}\right)$$

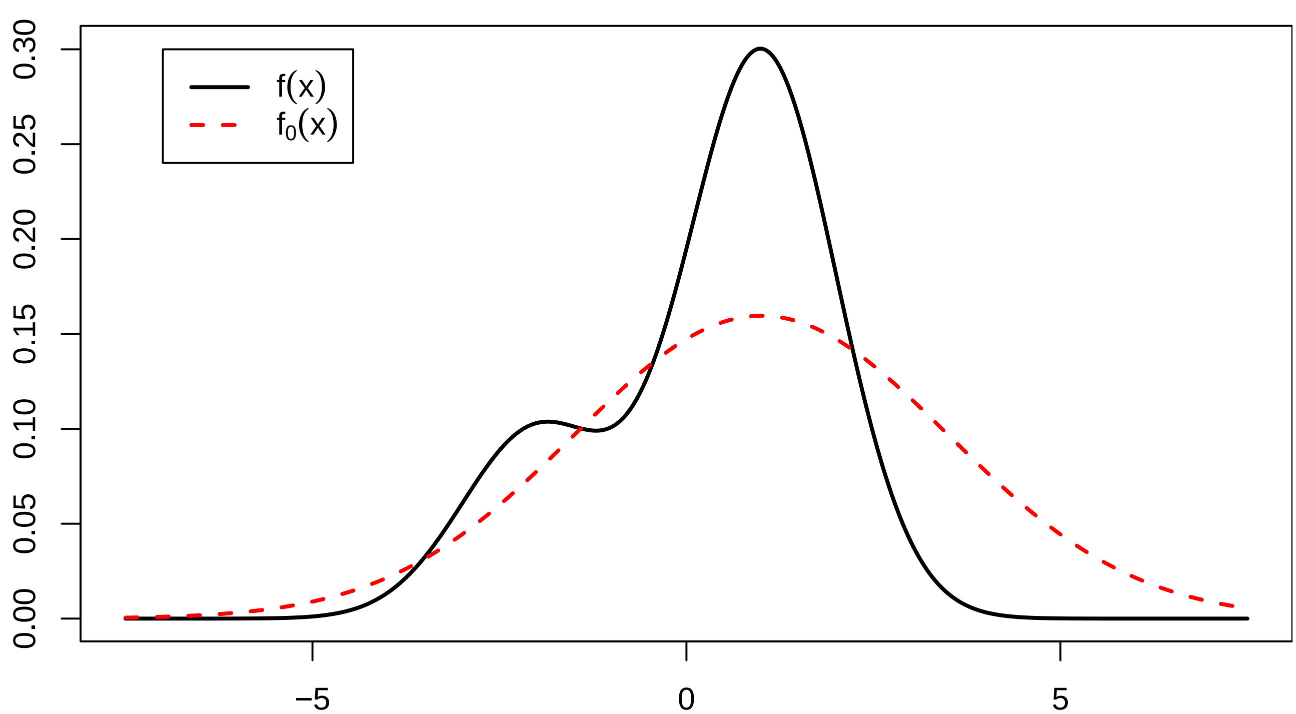
that is, a mixture of two Normal densities, where $\phi(.)$ is the standard Normal pdf. In the example below

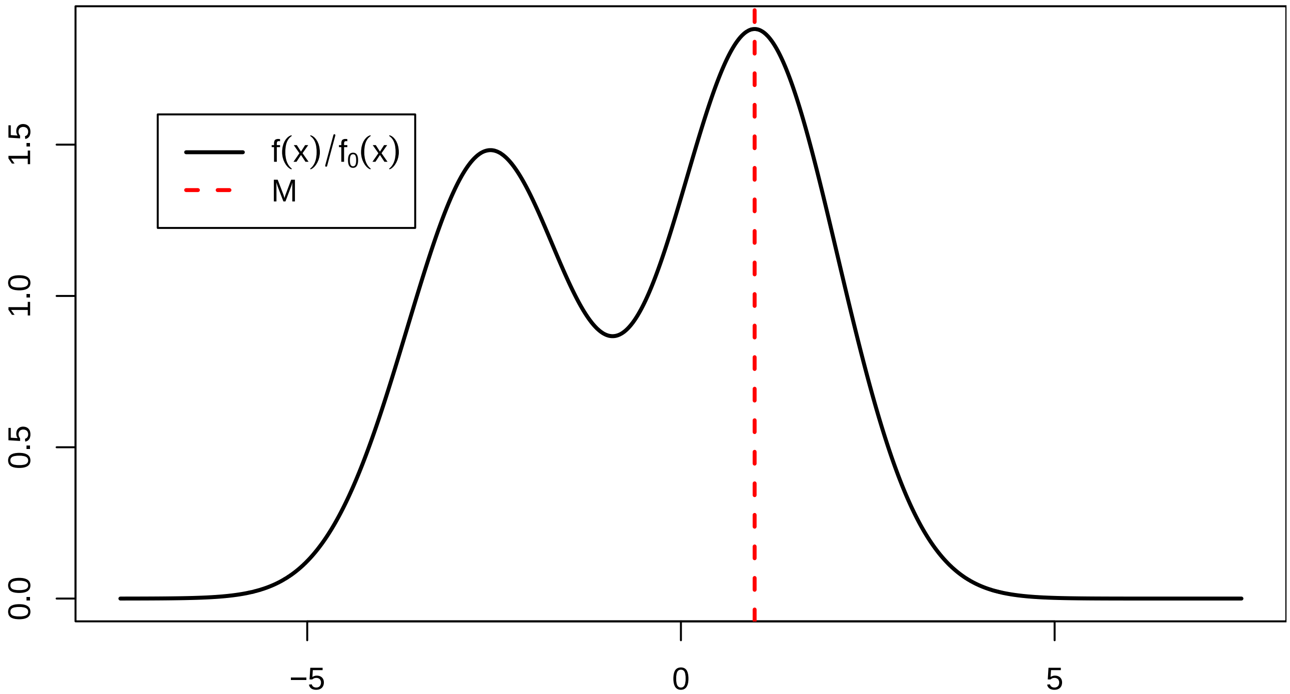
$$\mu_1 = -2 \qquad \sigma_1 = 1 \qquad \mu_2 = 1 \qquad \sigma_2 = 1.$$

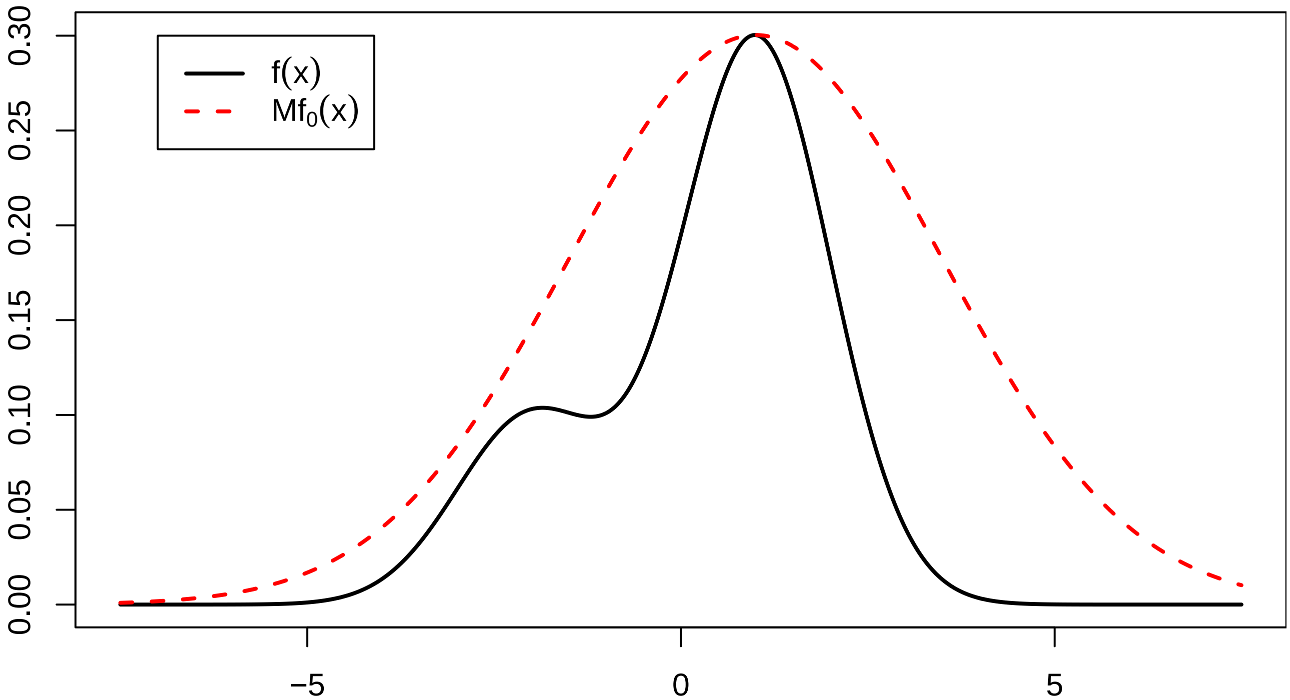
For $f_0(x)$ we propose to use the $Normal(1, 2.5^2)$ distribution.

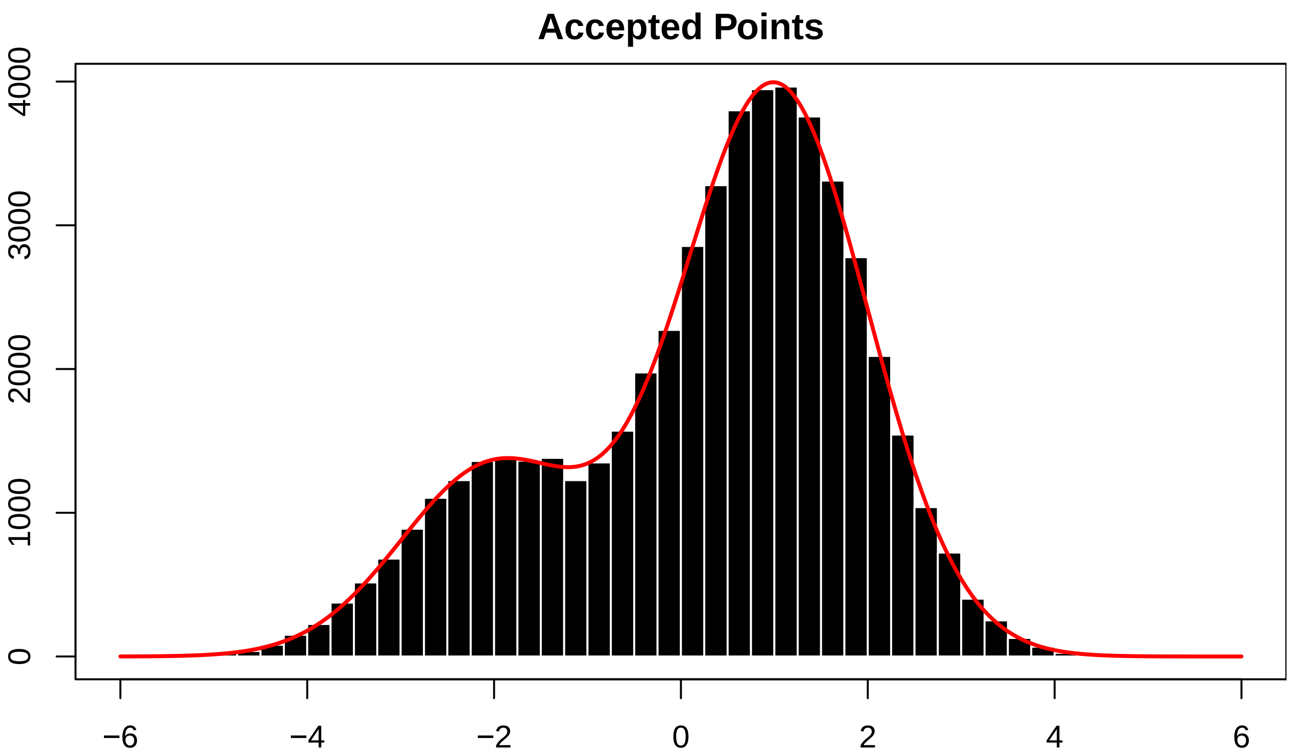
```
xv<-seq(-7.5,7.5,by=0.001)
fx<-0.25*dnorm(xv,-2,1)+0.75*dnorm(xv,1,1)
f0x<-dnorm(xv,1,2.5)
f.ratio<-function(xs){
        fxv<-0.25*dnorm(xs,-2,1)+0.75*dnorm(xs,1,1)
        f0xv<-dnorm(xs,1,2.5)
        return(fxv/f0xv)
}
par(mar=c(3,3,2,0))
plot(xv,fx,type="l",lwd=2)
lines(xv,f0x,col="red",lwd=2,lty=2)
legend(-7,0.3,c(expression(f(x)),expression(f[0](x))),col=c('black','red'),lty=c(1,2),lwd=c(2,2))
mval<-optim(c(0),fn=f.ratio,control=list(fnscale=-1),method="BFGS")
M<-max(fx/f0x)
```

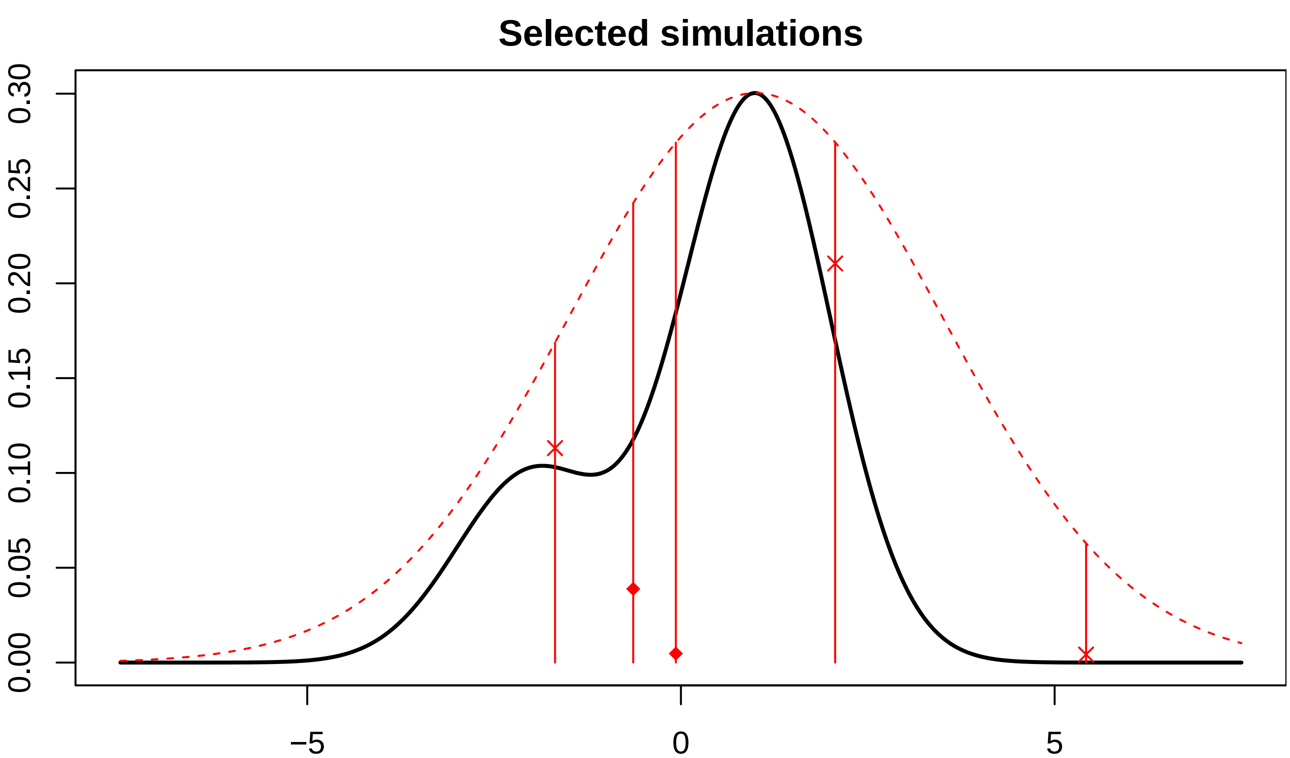
```
plot(xv,fx/f0x,type="l",xlab="x",ylab=expression(f(x)/f[0](x)),lwd=2)
abline(v=mval$par,col="red",lty=2,lwd=2)
legend(-7,1.6,c(expression(f(x)/f[0](x)),'M'),col=c('black','red'),lty=c(1,2),lwd=c(2,2))

plot(xv,fx,type="l",lwd=2)
lines(xv,mval$val*f0x,col="red",lwd=2,lty=2)
legend(-7,0.3,c(expression(f(x)),expression(M*f[0](x))),col=c('black','red'),lty=c(1,2),lwd=c(2,2))
M<-mval$val
N<-100000
U<-runif(N)
X<-rnorm(N,1,2.5)
pX<-(0.25*dnorm(X,-2,1)+0.75*dnorm(X,1,1))
p0X<-dnorm(X,1,2.5)
ival<-U<(pX/(M*p0X))
X.acc<-X[ival]
acc.rate<-sum(ival)/N

hist(X.acc,br=seq(-6,6,by=0.25),main="Accepted Points",col="black",border="white",xlab="x")
Xv<-seq(-6,6,by=0.01)
Yv<-(0.25*dnorm(Xv,-2,1)+0.75*dnorm(Xv,1,1))
lines(Xv,Yv*length(X.acc)*0.25,col="red",lwd=2);box()
plot(xv,fx,type="l",lwd=2,main="Selected simulations",ylab="Density",xlab="x")
lines(xv,mval$val*f0x,col="red",lty=2)
for(id in c(1000,7200,6000,5000,10000)){
    xh<-X[id]
    yh<-mval$val*dnorm(xh,1,2.5)
    lines(c(xh,xh),c(0,yh),col="red")
    u<-runif(1);acc<-4*(u>yh)+18*(u<yh)
    points(xh,u*yh,col="red",pch=acc,cex=1)
}
legend(-7,1.6,c('Accepted','Rejected'),col='red',pch=c(19,4))
```




f(x)/f0(x)
M
0.0
0.5
1.0
1.5
−5
0
5


f(x)
Mf0(x)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−5
0
5


Accepted Points
0
1000
2000
3000
4000
−6
−4
−2
0
2
4
6


Selected simulations
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−5
0
5

Antithetic Variables: The approach of antithetic variables exploits the fact that the average of two negatively correlated random variables can have a variance that is lower than the average variance of the variables. For example, if X_1 and X_2 have the same distribution, but are negatively correlated, then

$$\text{Var}\left[\frac{X_1 + X_2}{2}\right] = \frac{1}{4}\text{Var}[X_1] + \frac{1}{4}\text{Var}[X_2] + \frac{1}{4}\text{Cov}[X_1, X_2] = \frac{1}{2}\text{Var}[X_1] + \frac{1}{4}\text{Cov}[X_1, X_2] < \frac{1}{2}\text{Var}[X_1].$$

Applying this principle to Monte Carlo calculations, we try to construct samples $X_{1i}, \ldots, X_{1N}$ and $X_{2i}, \ldots, X_{2N}$ such that $(g(X_{1i}), g(X_{2i}))$ are negatively correlated so that

$$\frac{1}{2N}\sum_{i=1}^{N}(g(X_{1i}) + g(X_{2i}))$$

has a lower variance than the ordinary Monte Carlo estimator. Consider the following example: suppose we want to compute

$$\int_0^1 (x + x^2)\, dx = \frac{5}{6}$$

that is with $g(x) = x + x^2$, using Monte Carlo by sampling independently from a $Uniform(0, 1)$ distribution. Here

$$X_1 = X \qquad X_2 = 1 - X$$

have the **same** distribution, and are **negatively correlated**, and therefore the estimator

$$\frac{1}{2N}\sum_{i=1}^{N}(g(X_{1i}) + g(X_{2i}))$$

should have a lower variance than the estimator

$$\frac{1}{2N}\sum_{i=1}^{2N} g(X_i)$$

```
set.seed(64)
N<-5000
nreps<-1000
AV.ests<-matrix(0,nrow=nreps,ncol=2)
for(irep in 1:nreps){
    X<-runif(2*N)
    AV.ests[irep,1]<-mean(X+X^2)
    X1<-runif(N)
    X2<-1-X1
    AV.ests[irep,2]<-mean(X1+X1^2 + X2+X2^2)/2
}
(true.val<-5/6)

+ [1] 0.8333333

apply(AV.ests,2,mean)

+ [1] 0.8333084 0.8333483

cor(X1+X1^2,X2+X2^2)

+ [1] -0.9680439

N*apply(AV.ests,2,var)

+ [1] 0.16309497 0.00539412

var(AV.ests[,1])/var(AV.ests[,2])

+ [1] 30.23569
```

In this example there is a 30-fold decrease in variance of the estimator.