Variance reduction techniques

Ignacio Cascos

2019

Outline

4.1 Antithetic variables

Commonly simulation techniques are used to estimate a value θ that represents the mean of a rv $\theta = \mathbb{E}[X]$.

A sequence of observations of X denoted $X_1, X_2, \ldots, X_n$ is generated and θ is estimated as the average value $n^{-1}\sum_{i=1}^n X_i$.

If the estimator is unbiased, the Mean Square Error of the estimate is its variace

$$\text{MSE}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i - \theta\right)^2\right] = \frac{1}{n^2}\text{Var}\left[\sum_{i=1}^n X_i\right].$$

If the X_is are *independent*, the MSE is $\text{Var}[X]/n$.

Consider the situation at which we are simulating $n/2$ bivariate rvs (X_i, Y_i) all of them with the same distribution, the distribution of all the marginals X and Y being the same, and $\mathbb{E}[X] = \theta$, then $\text{Cov}(X_i, Y_i)$ is fixed for every i and possibly different from 0, while $\text{Cov}(X_i, X_j) = \text{Cov}(Y_i, Y_j) = 0$ otherwise.

$$\frac{1}{(n/2)^2}\text{Var}\left[\sum_{i=1}^{n/2}\frac{X_i+Y_i}{2}\right] = \frac{\text{Var}[X] + 2\text{Cov}(X,Y) + \text{Var}[Y]}{2n} = \frac{\text{Var}[X] + \text{Cov}(X,Y)}{n}.$$

If $\text{Cov}(X,Y) < 0$, then the variance of this estimator of θ is less than the variance of the estimator obtained out of n independent rvs.

If X is obtained as a transformation of k random numbers in $(0,1)$ as

$$X = h(U_1, U_2, \ldots, U_k),$$

where each $U_i \sim \text{U}(0,1)$ and independent one from the others, then $1 - U_i \sim \text{U}(0,1)$ and independent one from the others.

In conclusion,

$$Y = h(1-U_1, 1-U_2, \ldots, 1-U_k),$$

follows the distribution of X.

Further, under monotonicity assumptions on h, the rvs X and Y are negatively correlated.

Example

$$\int_0^1 e^x \,\mathrm{d}x = e - 1 = 1.718282$$

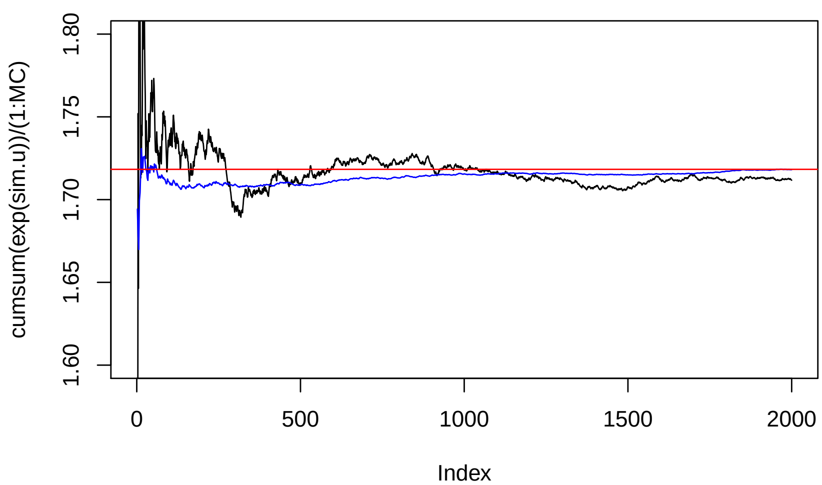
```
set.seed(1) ; MC <- 2000
sim.u <- runif(MC)
sim.exp <- exp(sim.u)
sim.ant <- (exp(sim.u[1:(MC/2)])+exp(1-sim.u[1:(MC/2)]))/2
sd(sim.exp)/sqrt(MC)
```

```
## [1] 0.01114951
```

```
sd(sim.ant)/sqrt(MC/2)
```

```
## [1] 0.001992196
```

```
plot(cumsum(exp(sim.u))/(1:MC),type="l",ylim=c(1.6,1.8))
bytwo <- seq(2,MC,by=2)
points(bytwo,cumsum(sim.ant)/(1:(MC/2)),type="l",col="blue")
abline(h=exp(1)-1,col="red")
```



- If the basic ingredient of our simulation algorithm is $U \sim \mathrm{U}(0,1)$, use the pair of negatively correlated and identically distributed rvs $U, 1-U$.
- If the basic ingredient of our simulation algorithm is $X \sim \mathrm{N}(\mu,\sigma)$, use use the pair of negatively correlated and identically distributed rvs $X, 2\mu - X$.
- If the basic ingredient of our simulation algorithm has cdf F, use the pair of negatively correlated and identically distributed rvs $F^{-1}(U), F^{-1}(1-U)$.

4.2 Control variates

Our goal is to estimate $\theta = \mathbb{E}[X]$, where X is a rv that we can simulate.

We can also simulate rv Y which is (negatively) correlated with X and whose mean is $\mathbb{E}[Y] = \mu_Y$.

For any constant c, the mean of $X + c(Y - \mu_y)$ is θ, and its variance is

$$\mathrm{Var}[X + c(Y - \mu_y)] = \mathrm{Var}[X] + c^2\mathrm{Var}[Y] + 2c\mathrm{Cov}(X, Y),$$

which is minimized for

$$c^* = \frac{|\mathrm{Cov}(X, Y)|}{\mathrm{Var}(Y)}$$

and the resulting variance is

$$\mathrm{Var}[X + c^*(Y - \mu_y)] = \mathrm{Var}[X] - \frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Var}(Y)}.$$

Example

$$\int_0^1 e^x \,\mathrm{d}x = e - 1 = 1.718282$$

We know $\int_0^1 (1 - x)\,\mathrm{d}x = 1/2$, and the rvs e^U and $1 - U$ are negatively correlated.

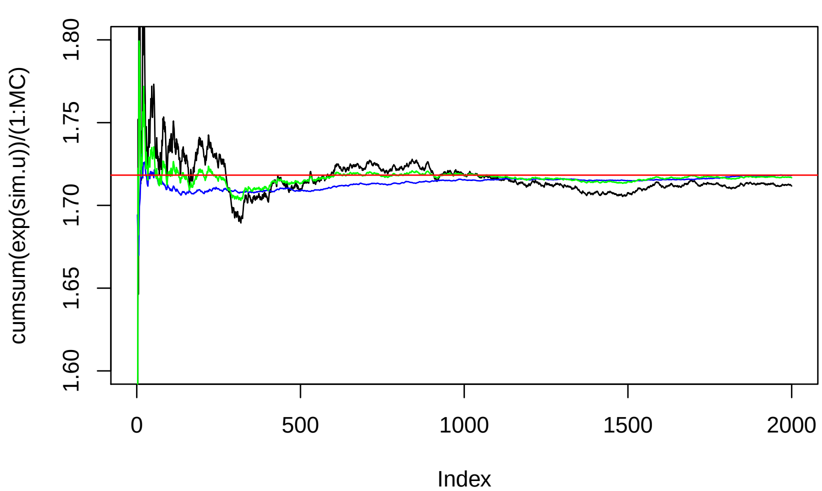
```
sim.cv <- exp(sim.u)+(1-sim.u-.5)
mean(sim.cv)
```

```
## [1] 1.716722
```

```
sd(sim.cv)/sqrt(MC)
```

```
## [1] 0.004738436
```

```
plot(cumsum(exp(sim.u))/(1:MC),type="l",ylim=c(1.6,1.8))
points(bytwo,cumsum(sim.ant)/(1:(MC/2)),type="l",col="blue")
points(cumsum(sim.cv)/(1:MC),type="l",col="green")
abline(h=exp(1)-1,col="red")
```



4.3 Stratified sampling

Our goal is to estimate $\theta = \mathbb{E}[X]$.

- There is a discrete variable Y with support $\{y_1, y_2, \ldots, y_k\}$ and known distribution $P(Y = y_i) = p_i$.
- For each $i = 1, \ldots, k$ we can simulate $X|Y = y_i$.

The **stratified sampling estimator** of θ is obtained after taking np_i observations of $X|Y = y_i$ for each $i = 1 \ldots, k$ (a total number of n observations is taken) and θ is estimated as $\sum_{i=1}^k \overline{X}_i p_i$, where $\overline{X}_i$ is the sample mean of X over the np_i instances for which $Y = y_i$.

Observe that (by the iterated expectation formula)

$$\mathbb{E}\left[\sum_{i=1}^k \overline{X}_i p_i\right] = \sum_{i=1}^k \mathbb{E}[\overline{X}_i] p_i = \sum_{i=1}^k \mathbb{E}[X|Y = y_i] p_i = \mathbb{E}[X]\,.$$

Variance of the stratified sampling estimator

$$\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^k \overline{X}_i p_i\right] &= \sum_{i=1}^k p_i^2 \mathrm{Var}[\overline{X}_i] \\
&= \sum_{i=1}^k p_i^2 \frac{1}{np_i} \mathrm{Var}[X|Y = y_i] \\
&= \frac{1}{n} \mathbb{E}[\mathrm{Var}[X|Y]].
\end{aligned}$$

Apply the conditional variance formula

$$\mathrm{Var}[X] = \mathbb{E}[\mathrm{Var}[X|Y]] + \mathrm{Var}[\mathbb{E}[X|Y]]$$

on $\overline{X}$ to check that

$$\mathrm{Var}[\overline{X}] - \mathrm{Var}\left[\sum_{i=1}^k \overline{X}_i p_i\right] = \frac{1}{n}\mathrm{Var}[\mathbb{E}[X|Y]] \geq 0$$

Example 1

```
p <- c(1728,1346,1869,1822,1056)/7821
sim.claim<-function(n) {
  claim<-runif(n,min=400,max=800)
  sim.u<-runif(n)
  bin1<-which(sim.u<p[1]) ; claim[bin1]<-runif(length(bin1),min=0,max=50)
  bin2<-which(sim.u>p[1] & sim.u<sum(p[1:2])) ; claim[bin2]<-runif(length(bin2),min=50,max=100)
  bin3<-which(sim.u>sum(p[1:2]) & sim.u<sum(p[1:3])) ; claim[bin3]<-runif(length(bin3),min=100,max=200)
  bin4<-which(sim.u>sum(p[1:3]) & sim.u<sum(p[1:4])) ; claim[bin4]<-runif(length(bin4),min=200,max=400)
  return(claim)
}
sclaim <- sim.claim(7821)
c(mean(sclaim),var(sclaim)/7821)
```

```
## [1] 203.902862   4.736561
```

```
sclaim <- list(runif(p[1]*7821,min=0,max=50),
               runif(p[2]*7821,min=50,max=100),
               runif(p[3]*7821,min=100,max=200),
```

```
                runif(p[4]*7821,min=200,max=400),
                runif(p[5]*7821,min=400,max=800))
(p[1]*mean(sclaim[[1]])+p[2]*mean(sclaim[[2]])+p[3]*mean(sclaim[[3]])
  +p[4]*mean(sclaim[[4]])+p[5]*mean(sclaim[[5]]))
```

```
## [1] 205.3114
```

```
(p[1]*var(sclaim[[1]])+p[2]*var(sclaim[[2]])+p[3]*var(sclaim[[3]])
  +p[4]*var(sclaim[[4]])+p[5]*var(sclaim[[5]]))/7821
```

```
## [1] 0.370737
```

```
var(sample(100000,x=c(25,75,150,300,600),prob=p,replace=T))/7821
```

```
## [1] 4.320348
```

Example 2

$$\int_0^1 e^x \, \mathrm{d}x = e - 1 = 1.718282$$

```
strata <- 10
sim.str <- exp(sim.u/strata+seq(0,.9,by=.1))
mstr <- vector(length=strata) ; vstr <- vector(length=strata)
for(i in 1:strata){
mstr[i]<-mean(sim.str[(1+(MC/strata)*(i-1)):((MC/strata)*i)])
vstr[i]<-var(sim.str[(1+(MC/strata)*(i-1)):((MC/strata)*i)])
}
mean(mstr)
```

```
## [1] 1.717646
```

```
mean(vstr)/MC
```

```
## [1] 0.0001220831
```

4.4 Importance sampling

Importance sampling is a *variance reduction technique*.

Assume we want to estimate p small, then the relative standard error of $\hat{p}_n$ is $\sigma(\hat{p}_n)/p = \sqrt{p(1-p)/n}p^{-1} \approx 1/\sqrt{np}$. If $p = 10^{-6}$, then the relarive standard error is $10^3/\sqrt{n}$ which is very large.

Consider rv X with dmf f, in order to estimate $\mathbb{E}[h(X)] = \int h(x)f(x)\,\mathrm{d}(x)$ use an instrumental density function g such that:

- $f(x) = 0$ whenever $g(x) = 0$

If Y stands for a rv with dmf g, then

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

where the weight function $\omega = f/g$.

By sampling from g and averaging $h(y_i)\omega(y_i)$ we obtain an unbiased estimation of $\mathbb{E}[h(Y)]$ whose variance is

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} h(Y_i)\omega(Y_i)\right] = \frac{1}{n}\left(\mathbb{E}[h^2(X)\omega(X)] - \mathbb{E}[h(X)]^2\right)$$

While when $g = f$, the variance is $(\mathbb{E}[h^2(X)] - \mathbb{E}[h(X)]^2)/n$. In order to reduce the variance we need that ω is small – equiv. g is large – for large values of h^2 (while it can be large – equiv. g small – at the values where h^2 is small).

```
1-pnorm(5)
```

```
## [1] 2.866516e-07
```

Example

```
set.seed(1) ; MC <- 10000
s.norm <- rnorm(MC)
c(mean(s.norm>5),var(s.norm>5)/MC)
```

```
## [1] 0 0
```

```
ss.norm <- s.norm+5
is <- (ss.norm>5)*exp((25-10*ss.norm)/2)
c(mean(is),var(is)/MC)
```

```
## [1] 2.799445e-07 4.565146e-17
```

Tilting

Density function

$$f_t(x) = \frac{e^{tx} f(x)}{M(t)}$$

is called *tilted density of* f, for $-\infty < t < \infty$.

Tilting is a natural way to obtain instrumental dmfs for importance sampling.

If $X \sim \mathrm{N}(\mu, \sigma)$, by tilting, we obtain $\mathrm{N}(\mu + t\sigma^2, \sigma)$.