Project 2

Computer Intensive Statistical Methods

Erling og Christian

03 02 2022

Contents

Caption

Problem C: Monte Carlo integration and variance reduction

Here we use Monte Carlo integration to estimate $\theta = P(X > 4)$ when $X \sim N(0, 1)$.

problem introduction. is it needed? thought it looked nice\ In this problem we consider Monte Carlo integration and the importance sampling for variance reduction.

1

Let $h(X) = I(X > 4)$, where I is the indicator function, such that

$$
\begin{aligned}
E[h(X)] &= \int_{-\infty}^{\infty} h(x) f(x) dx \\
&= \int_{-\infty}^{\infty} I(x > 4) f(x) dx \\
&= P(X > 4) \\
&= \theta.
\end{aligned}
$$

Then, a Monte Carlo Estimate of θ is given by

$$\hat{\theta}_{MC} = \frac{1}{n} \sum_{i=1}^{n} h(x_i).$$

Drawing $n = 100000$ samples from the standard distribution to estimate θ by MC.

```
generate_from_normal <- function(n) {
    u1 <- runif(n)
    u2 <- runif(n)
    r <- sqrt(-2 * log(u1))
```

```
    x_1 <- 2 * pi * u2
    x = r * cos(x_1)
    return(x)
}
```

Now we will find a $1-\alpha$ confidence for θ based on our sample set. First we need the expected value of the Monte Carlo estimator

$$
\begin{aligned}
E[\hat{\theta}] &= E[\frac{1}{n}\sum_{i=1}^{n} h(x_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\theta \\
&= \theta,
\end{aligned}
$$

and it's variance

$$
\begin{aligned}
Var(\hat{\theta}) &= Var(\frac{1}{n}\sum_{i=1}^{n} h(x_i)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(h(x_i)) \\
&= \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(h(x_i)-\hat{\theta})^2.
\end{aligned}
$$

As for the confidence interval we have the statistic

$$T_{MC} = \frac{\hat{\theta}_{MC}-\theta}{\sqrt{\hat{Var}(\hat{\theta}_{MC})}} \sim t_{n-1}$$

Why does the sample variance become so small?

```
set.seed(321)
n = 1e+05
x = generate_from_normal(n)
h = x > 4
MCest = mean(h)
# test
x2 = rnorm(n)
h2 = x2 > 4
theta2 = mean(h2)
theta = pnorm(4, lower.tail = F)
round(c(rnorm = theta2, inversion = MCest, True = theta), 7)
```

```
##     rnorm inversion      True
##  3.00e-05  5.00e-05  3.17e-05
```

```
# svMC = var(h) # Sample variance
svMC = sum((h - MCest)^2)/(n * (n - 1))  # Sample Variance
alpha = 0.05
t = qt(alpha/2, n - 1, lower.tail = F)  # (1-alpha) significance
lwrUpr = sqrt(svMC) * t  # lower and upper deviation from mean
ciMC = MCest + c(-lwrUpr, lwrUpr)
resultMC <- c(estimator = MCest, Confint = ciMC, Var = svMC)
resultMC
```

```
##    estimator     Confint1     Confint2          Var
## 5.000000e-05 6.174219e-06 9.382578e-05 4.999800e-10
```

```r
theta
```

```
## [1] 3.167124e-05
```

```r
# Below is a sanity check plot(density(x2), main = 'Comparing built in and
# inversion normal') lines(density(x), lty=2, col='red')
# legend('topright',legend = c('rnorm', 'Inversion'), lty=c(1,2), col =
# c('black', 'red'))
```

Something wrong with the variance?

2

Here we will use importance sampling to try to reduce the variance. The proposal distribution is

$$g(x) = \begin{cases} cxe^{-x^2/2} & , \ x > 4 \\ 0 & , \ \text{otherwise}, \end{cases}$$

where c is a normalizing constant.

Proposing to write importance sampling stuff here:

Now, let $x_1, ..., x_n \overset{i.i.d}{\sim} g(x)$ and let the weights $w(x_i) = w_i = f(x_i)/g(x_i) = f_i/g_i$. (where w_i, f_i and g_i are function evaluations at x_i). Then, the importance sampling estimator of θ is

$$\hat{\theta}_{IS} = \frac{\sum_{i=1}^n h_i w_i}{n}.$$

Since we will sample from g we need to find it's cdf

$$\begin{aligned} G(x) &= \int_4^x cye^{-y^2/2} dy \\ &= \int_8^{x^2/2} ce^{-u} du \\ &= \left[-ce^{-u}\right]_8^{x^2/2} \\ &= c(e^{-8} - e^{-x^2/2}). \end{aligned}$$

Since g is a distribution, $\int_4^\infty g(x)dx = 1$. Thus, we find c by considering

$$\begin{aligned} c(e^{-8} - e^{-x^2/2})\Big|_{x=\infty} &= 1 \\ c &= e^8. \end{aligned}$$

Furthermore, $G(x) = 1 - e^{8-x^2/2}$.

Inversion method for sampling: \ Then, by inversion method, we can sample from g by solving $U = G(x) \sim Unif(0,1)$ for x. We get

$$\begin{aligned} U &= 1 - e^{8-x^2/2} \\ -2ln(1-U) &= x^2 - 8 \\ x &= \sqrt{8 - 2ln(1-U)}, \end{aligned}$$

that is, inserting randomly selected $U \sim Unif(0,1)$ admits samples from $X \sim g$. Should we include this short chunk here or at the end?

```
# expSampler <- function(n){ u = runif(n) return(sqrt(16-2*log(1-u))) } gx <-
# expSampler(n)
```

We also need the expected value and sample variance of the importance estimator for the $(1-\alpha)$ confidence interval, so

or: For the (1-α) confidence interval we also need the expected value and the sample variance of the importance estimator hatθ.

is it really integral when x seem do be discrete, x_i?

$$
\begin{aligned}
E[\hat{\theta}_{IS}] &= E\left[\frac{\sum_{i=1}^{n} h_i w_i}{n}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty h_i \frac{f_i}{g_i} g_i dx \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty h_i f_i dx \\
&= \frac{1}{n}\sum_{i=1}^{n} E[h_i] \\
&= \frac{1}{n} n\theta \\
&= \theta,
\end{aligned}
$$

and the sample variance More steps in calculation?

$$
\begin{aligned}
Var(\hat{\theta}_{IS}) &= Var\left(\frac{\sum_{i=1}^{n} h_i w_i}{n}\right) \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(h_i w_i - \sum_{i=1}^{n}\frac{h_i w_i}{n}\right)^2 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(h_i w_i - \hat{\theta}_{IS}\right)^2
\end{aligned}
$$

```
expSampler <- function(n) {
    # Samples from prop fnc g
    u = runif(n)
    return(sqrt(16 - 2 * log(1 - u)))
}

w <- function(x) {
    f = dnorm(x)
    g = ifelse(x > 4, x * exp(8 - 0.5 * x^2), 0)
    return(f/g)
}

gx = expSampler(n)  # Sample from proposal
gh = (gx > 4) * 1
```

```r
ISest = mean(gh * w(gx))
svIS = sum((gh * w(gx) - ISest)^2)/(n * (n - 1))
ISconfint = ISest + c(-t * sqrt(svIS), t * svIS)
# Results
resultIS <- c(ISestimate = ISest, confint = ISconfint, var = svIS)
results <- rbind(MC = resultMC, IS = resultIS)
results
```

```
##       estimator     Confint1     Confint2          Var
## MC 5.000000e-05 6.174219e-06 9.382578e-05 4.999800e-10
## IS 3.167425e-05 3.166461e-05 3.167425e-05 2.419449e-17
```

```r
theta
```

```
## [1] 3.167124e-05
```

Here we see that importance sampling has reduced the variance by a factor of $Var(\hat{\theta}_{MC})/Var(\hat{\theta}_{IS})$ = `svMC/svIS` $= 2.0665032 \times 10^7$. Also, the importance sample estimator is much closer to the real value.

Not entirely sure how to compute how many more samples we would need, since the error influences.

3

Now we will combine the importance sampling with the use of antithetic variates. We start by modifying the sample generator for g in task $C.2$ so that it produces n pairs, $X = x_i$ and $Y = y_i$, of antithetic variates by using u_i and $1 - u_i$ as input to G^{-1}, respectively.

```r
expSamplerAnti <- function(n, u = NA) {
    u = runif(n)
    return(data.frame(x = sqrt(16 - 2 * log(1 - u)), y = sqrt(16 - 2 * log(u))))
}
# xy <- expSamplerAnti(n) gx <- ifelse(xy$x>4, xy$x*exp(8-0.5*xy$x^2),0) gy <-
# ifelse(xy$y>4, xy$y*exp(8-0.5*xy$y^2),0) plot(xy$x,gx) points(xy$y,gy,
# col='cyan3',cex=1)
```

Then, the importance sample estimates for each of the pairs are

$$\hat{\theta}_X = \frac{1}{n}\sum_{i=1}^{n} h(x_i)w(x_i),$$
$$\hat{\theta}_Y = \frac{1}{n}\sum_{i=1}^{n} h(y_i)w(y_i),$$

and the antithetic sample estimator is

$$\hat{\theta}_A = \frac{\hat{\theta}_X + \hat{\theta}_Y}{2}.$$

We also need the expected value

$$\begin{aligned}
E[\hat{\theta}_{AS}] &= E\left[\frac{\hat{\theta}_X + \hat{\theta}_Y}{2}\right] \\
&= \frac{1}{2n}\sum_{i=1}^{n}(E[h(x_i)w(x_i)] + E[h(y_i)w(y_i)]) \\
&\overset{*}{=} \frac{1}{2n}\sum_{i=1}^{n} 2\theta \\
&= \theta,
\end{aligned}$$

where $*$ since the proposal distribution evaluations cancel in each expectation. not true or not needed?. The variance of the estimator is

$$\begin{aligned}
Var(\hat{\theta}_{AS}) &= \frac{1}{4}(Var(\hat{\theta}_X) + Var(\hat{\theta}_Y) + 2Cov(\hat{\theta}_X, \hat{\theta}_Y)) \\
\Downarrow \quad & Var(\hat{\theta}_X) = Var(\hat{\theta}_Y) \\
&= \frac{(1+\rho_{XY})S^2_{XY}}{2n},
\end{aligned}$$

where $\rho_{XY} = Cov(\hat{\theta}_X, \hat{\theta}_Y)$ and S^2_{XY} is the sample variance of either estimator $\hat{\theta}_X$ or $\hat{\theta}_Y$.

```
set.seed(321)
n = 50000
xy = expSamplerAnti(n)
hxy = (xy > 4) * 1
hwx = hxy[, "x"] * w(xy$x)
hwy = hxy[, "y"] * w(xy$y)
hwxy = (hwx + hwy)/2
ASest = mean(hwxy)
var(hwxy)
```

```
## [1] 2.851883e-13
```

```
svAS = (var(hwx) + var(hwy) + 2 * cov(hwx, hwy))/4
lwrUprAS = c(-t, t) * sqrt(svAS)
confintAS = ASest + lwrUprAS
resultAS <- c(ASest, confintAS, svAS)
rbind(results, AS = resultAS)
```

```
##       estimator     Confint1     Confint2          Var
## MC 5.000000e-05 6.174219e-06 9.382578e-05 4.999800e-10
## IS 3.167425e-05 3.166461e-05 3.167425e-05 2.419449e-17
## AS 3.167262e-05 3.062593e-05 3.271931e-05 2.851883e-13
```