homework 5

Topics in Statistics 1, Jieun Shin

2021.10.28

```
library(dplyr)
```

5.7

$\mathbb{V}(U) = \mathbb{E}[\mathbb{V}(U|V)] + \mathbb{V}[\mathbb{E}(U|V)]$ 에서 오른쪽 변의 첫 번째 항을 정리하면

$$\begin{aligned}
\mathbb{V}(U|V) &= \mathbb{E}(U^2|V) - \mathbb{E}(U|V)^2 \\
\mathbb{E}[\mathbb{V}(U|V)] &= \mathbb{E}[\mathbb{E}(U^2|V)] - \mathbb{E}[\mathbb{E}(U|V)^2] \quad (\text{양 변에 기댓값을 씌움}) \\
&= \mathbb{E}(U^2) - \mathbb{E}[\mathbb{E}(U|V)^2]
\end{aligned}$$

이고, 두 번째 항을 정리하면

$$\begin{aligned}
\mathbb{V}[\mathbb{E}(U|V)] &= \mathbb{E}[\mathbb{E}(U|V)^2] - \mathbb{E}[\mathbb{E}(U|V)]^2 \quad (\because \mathbb{V}(V) = \mathbb{E}(V^2) - \mathbb{E}(V)^2) \\
&= \mathbb{E}[\mathbb{E}(U|V)^2] - \mathbb{E}(U)^2
\end{aligned}$$

이다. 따라서 두 식을 더하면

$$\begin{aligned}
\mathbb{E}[\mathbb{V}(U|V)] + \mathbb{E}[\mathbb{E}(U|V)] &= \mathbb{E}(U^2) - \mathbb{E}[\mathbb{E}(U|V)^2] + \mathbb{E}[\mathbb{E}(U|V)^2] - \mathbb{E}(U)^2 \\
&= \mathbb{E}(U^2) - \mathbb{E}(U)^2 \\
&= \mathbb{V}(U)
\end{aligned}$$

으로 등식을 만족한다.

5.8

(a) 확률변수 R은 기하분포 $\mathrm{G}(p)$를 따르고, $X_1, X_2, \cdots \sim \mathrm{Exp}(\lambda)$의 R개 합을 새로운 확률변수 $Y = \sum_{i=1}^{R} X_i$는 $\mathrm{Gamma}(R, \lambda)$를 따른다. Y의 cdf를 구하면 다음과 같다.

$$\begin{aligned}
F_Y(y) = \mathbb{P}(Y \le y) &= \mathbb{P}(Y \le y|R = r)\mathbb{P}(R = r) \\
&= \sum_{r=0}^{\infty} \int_{t=0}^{y} \frac{\lambda^r}{\Gamma(r)} t^{r-1} e^{-\lambda t} dt p(1-p)^{r-1} \\
&= \int_{t=0}^{y} \lambda p e^{\lambda t} \sum_{r=0}^{\infty} \frac{\{\lambda t(1-p)\}^{(r-1)}}{(r-1)!} dt \\
&= \int_{t=0}^{y} \lambda p e^{\lambda p t} dt \\
&= 1 - e^{\lambda p y}
\end{aligned}$$

y에 대해 미분하면 Y의 분포는 $f_Y(y) = \lambda p e^{-\lambda p y}$ 이다. 따라서 $Y = S_R \sim \mathrm{Exp}(\lambda p)$ 을 따른다.

(b) $\lambda = 1, p = 1/10, N = 1000$에 대해 $\mathbb{P}(S_R > 10)$의 CMC 추정치는 $\frac{1}{1000}\sum_{i=1}^{1000} I(S_R > 10)$이다. 아래 코드에 의해 추정치는 약 0.393이고 분산은 약 0.239이다.

```
set.seed(123)
N = 1000

# (b)
lambda = 1
p = 0.1
x = rexp(N, lambda * p)

mean(x > 10)   # CMC
```

```
## [1] 0.393
```

```
var(x > 10)   # CMC
```

```
## [1] 0.2387898
```

(c) (b)를 반복하여 조건부 몬테카를로 추정치를 구한 결과이다. 추정치는 약 0.36, 분산은 약 0.231으로 CMC의 분산보다 작은것을 확인할 수 있다.

```
# (c)
g = rgeom(N, p) + 1
s = sapply(1:N, function(x) sum(rexp(g[x], lambda)))
mean(s > 10)
```

```
## [1] 0.36
```

```
var(s > 10)
```

```
## [1] 0.2306306
```

5.9

문제 5.8에 이어서 $p = 0.25, \lambda = 1$에 대해 $\mathbb{P}(S_R > 10)$를 층화추출법으로 추정해보자. 먼저 층을 $\{R = 1\}, \ldots, \{R = 7\}, \{R > 7\}$의 8개의 층으로 나누어 각 층의 확률을 $R \sim \mathrm{G}(p)$에 대하여 $\mathbb{P}(R = 1), \ldots, \mathbb{P}(R = 7), \mathbb{P}(R > 7)$로 계산할 수 있으며, 계산한 각 층의 확률 p_i는 $0.2500, 0.1875, \ldots, 0.1335$이다. 각 층의 크기 $N_i, i = 1, \ldots, 8$은 \$2500, ..., 1335 \$이고, N_i과 p_i를 통해 추정치와 추정치의 분산을 구할 수 있다. 추정치는 $\hat{l}^s = \sum_{i=1}^{8} p_i \frac{1}{N_i} \sum_{j=1}^{N_j} I(X_{ij} > 10) = 0.0846$, 추정치의 분산은 $j = 1, \ldots, N_i$에 대해 $Var(\hat{l}^s) = \sum_{i=1}^{8} \frac{p_i^2}{N_i} Var(I(X_{ij} > 10)|Y = i) \approx 4.621267e - 06$ 이다.

```
set.seed(123)
p = 0.25
lambda = 1
N = 10000

# sampling
g = rgeom(N, p) + 1
```

```
s = sapply(1:N, function(x) sum(rexp(g[x], lambda)))

# calculate the n_i and Pr(y=i) for i=1, 2, ..., 8
gg = ifelse(g > 7, 8, g)
pi = dgeom(0:6, p)
pi[8] = 1 - sum(pi)

pi  # 각 층의 확률
```

```
## [1] 0.25000000 0.18750000 0.14062500 0.10546875 0.07910156 0.05932617 0.04449463
## [8] 0.13348389
```

```
sigma = sapply(1:length(pi), function(x){var(s[which(gg == x)] > 10)})

# find optimal Ni
Ni = pi * N

# mean estimation
m = sapply(1:length(pi), function(x) sum(s[which(gg == x)] > 10) / Ni[x])

# stratified sampling estimator
sum(pi * m)
```

```
## [1] 0.0846
```

```
# variance of stratified sampling estimator
sum(pi^2 * sigma / Ni)
```

```
## [1] 4.621267e-06
```

5.15

$\mathbb{P}(Z > 4)$를 중요도 샘플링을 사용하여 추정해보자. 먼저 $l = \mathbb{P}(Z > 4) = \mathbb{E}I(Z > 4)$이므로 $H(X) = I(X > 4)$이고, 따라서 l의 추정치는 $\hat{l} = \frac{1}{N}\sum_{i=1}^{N} I(x_i > 4)$이다.

중요도 샘플링을 위해 확률변수 X를 중요도 함수 $g(x) = e^{-(x-4)}, \quad x \geq 4$하자. 그러면 $f \sim N(0,1)$이므로 중요도 샘플링을 이용한 추정치는 $\hat{l}^s = \frac{1}{N}\sum_{i=1}^{N} I(x_i > 4)\frac{f(x_i)}{g(x_i)} = \frac{1}{N}\sum_{i=1}^{N} I(x_i > 4)\frac{1}{\sqrt{2\pi}}e^{-x^2/2+x-4}$로 구할 수 있다.

이렇게 계산한 추정치는 3.163248e-05이며 분산은 1.461636e-09 이다. 추정치는 참값 $\int_4^\infty \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = 3.167124e-05$으로 추정치가 참값과 매우 비슷한 값을 가진다.

```
rm(list = ls())
set.seed(123)
N = 1e+7

rg = function(N) 4 -log(runif(N))
fg = function(x) exp( -(x^2)/2 + x - 4) / sqrt(pi * 2)

x = rg(N)
mean(fg(x))
```

```
## [1] 3.166363e-05
```

```r
var(fg(x))
```

```
## [1] 1.461679e-09
```

```r
# true
integrate(dnorm, 4, 10)
```

```
## 3.167124e-05 with absolute error < 1.1e-11
```

5.17

먼저 $X_k \sim f(x; \boldsymbol{u}), \quad k = 1, \ldots, 5$ 라 하면 CE update를 한 $\boldsymbol{v}^*$ 의 각 v_i^* 를 다음과 같은 식에 의해 찾을 수 있다.

$$v_i^* = \frac{\sum_{k=1}^N H(x_k) x_{ki}}{\sum_{k=1}^N H(x_k)},$$

코드를 통해 $\boldsymbol{v}^* = (0.88, 0.88, 1.99, 0.45, 0.50)$ 의 값이 나왔다.

$f(x; \boldsymbol{u})$ 를 평균이 $\boldsymbol{u} = (u_1, \ldots, u_5) = (1, 1, 0.5, 2, 1.5)$ 인 지수분포를 따른다고 하자. 그리고 $f(x; \boldsymbol{v}^*)$ 를 평균이 $\boldsymbol{v}^* = (v_1^*, \ldots, v_5^*)$ 인 지수분포를 따른다 하자 (즉 f 와 g 는 같은 분포족에 속한다.). 그러면 $\frac{f(x;\boldsymbol{u})}{f(x;\boldsymbol{v}^*)}$ 은 다음과 같이 정리할 수 있다.

$$\frac{f(x; \boldsymbol{u})}{f(x; \boldsymbol{v}^*)} = \frac{\prod_{i=1}^5 \frac{1}{u_i} e^{-x_i/u_i}}{\prod_{i=1}^5 \frac{1}{v_i^*} e^{-x_i/v_i^*}} = \exp\left(-\sum_{i=1}^5 x_i \left(\frac{1}{u_i} - \frac{1}{v_i^*} \right) \right) \prod_{i=1}^5 \frac{v_i^*}{u_i}$$

중요도 샘플링을 이용한 추정은 $\hat{l} = \frac{1}{N} \sum_{i=1}^N H(x_i) \frac{f(x;\boldsymbol{u})}{f(x;\boldsymbol{v}^*)}$ 으로 추정할 수 있으며, 그 결과 추정치는 약 1.52, 분산은 약 0.49이다.

```r
set.seed(123)
H = function(p){
  x = rexp(5, p)

  h = min(x[1] + x[4], x[1] + x[3] + x[5], x[2] + x[3] + x[4], x[2] + x[5])
  w = h * x

  out = list()
  out$w = w
  out$h = h

  return(out)
  }

N = 1000
prob = 1 / c(1, 1, 0.5, 2, 1.5)

h = replicate(N, H(prob)$h)
w = replicate(N, H(prob)$w)

# find optimal v
optv = 1 / (rowSums(w)/sum(h))
optv
```

```
## [1] 0.8800820 0.8788947 1.9941091 0.4510440 0.5011815
```

```r
# estimate H(x) using importance sampling
fg = function(x, u, v) exp( -sum(x * (v - u) /(u * v) )) * prod(v / u)

X = replicate(N, rexp(5, optv))
W = sapply(1:N, function(x) fg(X[,x], 1/prob, 1/optv))

hh = function(x) min(x[1] + x[4], x[1] + x[3] + x[5], x[2] + x[3] + x[4], x[2] + x[5])

hval = sapply(1:N, function(x) hh(X[,x]))

mean(hval * W)
```

```
## [1] 1.524295
```

```r
var(hval * W)
```

```
## [1] 0.4888715
```

5.18

f가 지수족 분포이므로 $f(\boldsymbol{x};\boldsymbol{\theta}) = h(\boldsymbol{x})\exp(\eta(\boldsymbol{\theta})T(\boldsymbol{x}))c(\boldsymbol{\theta})$와 같이 일반화하여 표현할 수 있다. $f(\boldsymbol{x};\boldsymbol{\theta})$에 로그를 취하면

$$\log f(\boldsymbol{x};\boldsymbol{\theta}) = \log h(\boldsymbol{x}) + \eta(\boldsymbol{\theta})T(\boldsymbol{x}) + \log c(\boldsymbol{\theta})$$

이고, 이어서 $\boldsymbol{\theta}$에 대해 미분하면

$$\nabla \log f(\boldsymbol{x};\boldsymbol{\theta}) = \nabla\eta(\boldsymbol{\theta})T(\boldsymbol{x}) + \nabla \log c(\boldsymbol{\theta})$$

이다.

식 (5.112)의 경우 $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}$ 이므로 $\nabla\eta(\boldsymbol{\theta}) = 1$이다. 따라서 위 식을 정리하면

$$\nabla \log f(\boldsymbol{x};\boldsymbol{\theta}) = T(\boldsymbol{x}) + \frac{\nabla c(\boldsymbol{\theta})}{c(\boldsymbol{\theta})}$$

이다.

따라서 f가 지수족 분포인 경우 $\boldsymbol{u} = \boldsymbol{\theta}_0, \boldsymbol{v} = \boldsymbol{\theta}$에 대하여 등식

$$\mathbb{E}_{\boldsymbol{u}}[H(\boldsymbol{X})\nabla \log f(\boldsymbol{X};\boldsymbol{v})] = \mathbb{E}_{\boldsymbol{\theta}_0}\left[H(\boldsymbol{X})\left(T(\boldsymbol{x}) + \frac{\nabla c(\boldsymbol{\theta})}{c(\boldsymbol{\theta})}\right)\right]$$

이 성립한다.