Lecture 5: Monte Carlo 积分与方差减少技术

张伟平

Sunday 27$^{\text{th}}$ September, 2009

Contents

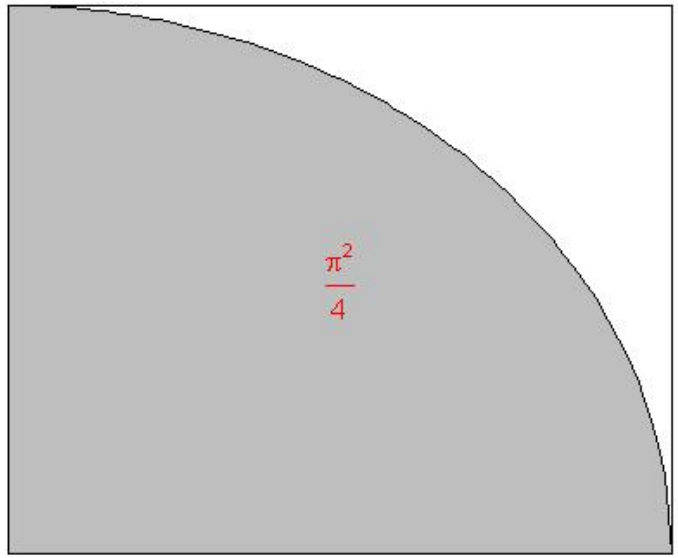
Chapter 1

Monte Carlo Integration and Variance Reduction

蒙特卡罗(Monte Carlo)积分是一种基于随机抽样的统计方法. 蒙特卡罗方法其实 也只是对一种思想的泛称, 只要在解决问题时, 利用产生大量随机样本，然后对 这些样本结果进行概率分析, 从而来预测结果的方法, 都可以称为蒙特卡罗方法.

比如要求圆周率π的值, 最著名的就是中学时学过的"割圆法"(刘徽(魏晋, 3.1416), 祖冲之(南北朝, 3.1415926$<\pi<$3.1415927)). 现在也可以使用蒙特卡罗积分方法: 由概率论知

$$\text{若 } r.v.\ X, Y\ i.i.d\ U(0,1),\ \text{则 } P(X^2+Y^2\le 1)=\frac{\pi}{4}$$



因此, $\pi = 4P(X^2 + Y^2 \leq 1) \approx 4\#\{x^2 + y^2 \leq 1\}/n$.

- $n = 1000$, $\hat{\pi} = 3.168$.
- $n = 100,000$, $\hat{\pi} = 3.14312$.
- $n = 10^7$, $\hat{\pi} = 3.141356$.

1.1 Monte Carlo Integration

假设g是一个可积函数, 我们要计算$\int_a^b g(x)dx$. 回忆在概率论中, 若随机变量X的密度为$f(x)$, 则随机变量$Y = g(X)$的期望为

$$Eg(X) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

如果从X的分布中产生了一些随机数, 则$Eg(X)$的无偏估计就是相应的样本平均值.

1.1.1 Simple Monte Carlo estimator

考虑估计$\theta = \int_0^1 g(x)dx$. 若$X_1, \cdots, X_m$为均匀分布$U(0,1)$总抽取的样本, 则由强大数律知

$$\hat{\theta} = \overline{g_m(X)} = \frac{1}{m}\sum_{i=1}^{m} g(X_i)$$

以概率1收敛到期望$Eg(X)$. 因此$\int_0^1 g(x)dx$的简单的Monte Carlo 估计量为$\overline{g_m(X)}$.

例1: (简单的Monte Carlo 积分) 计算

$$\theta = \int_0^1 e^{-x} dx$$

的简单Monte Carlo估计以及与积分值相比较.

↑Example

```
m <- 10000
x <- runif(m)
theta.hat <- mean(exp(-x))
print(theta.hat)
print(1 - exp(-1))
#[1] 0.6355289
#[1] 0.6321206
```

↓Example

估计为$\hat{\theta} \doteq 0.6355$, 而积分值为$\theta = 1 - e^{-1} \doteq 0.6321$.

若要计算$\int_a^b g(x)dx$, 此处$a < b$为有限数. 则作一积分变量代换使得积分限从0到1. 即作变换$y = (x - a)/(b - a)$, 因此

$$\int_a^b g(x)dx = \int_0^1 g(y(b-a)+a)(b-a)dy$$

另外一种做法就是利用均匀分布$U(a,b)$, 即

$$\int_a^b g(x)dx = (b-a)\int_a^b g(x)\frac{1}{b-a}dx$$

例2: 简单Monte Carlo 积分(续) 计算

$$\theta = \int_2^4 e^{-x}dx$$

的Monte Carlo估计并和积分值相比较.

↑Example

```
m <- 10000
x <- runif(m, min=2, max=4)
theta.hat <- mean(exp(-x)) * 2
print(theta.hat)
print(exp(-2) - exp(-4))
#[1] 0.1172158
#[1] 0.1170196
```

即, 估计的值为$\hat{\theta} \doteq 0.1172$, 而真值为$\theta = e^{-2} - e^{-4} \doteq 0.1170$.

总结一下, 积分$\int_a^b g(x)dx$的简单Monte Carlo估计方法为

1. 从均匀分布$U(a,b)$中产生$i.i.d$样本$X_1, \cdots, X_m$;
2. 计算$\overline{g_m(X)} = \frac{1}{m} g(X_i)$
3. $\hat{\theta} = (b-a)\overline{(g_m(X))}$.

例3: Monte Carlo积分, 无穷积分 比如使用Monte Carlo积分方法计算标准正态的分布函数

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

首先我们不能直接使用以前的方法(因为积分区间无界), 但是我们可以将此问题分为两种情形: $x > 0$和$x \leq 0$. 若 $x > 0$, 注意到对标准正态分布, 积分

区间可以分为$(-\infty, 0)$和$(0, x)$,因此只需要计算积分$\theta = \int_0^x e^{-\frac{t^2}{2}} dt$ 即可. 故可以使用之前的方法. 但是需要从均匀分布$U(0, x)$中抽取样本, 若x发生变化, 则均匀分布也就变化了. 若要求从$U(0, 1)$中抽样, 则可以作变换$y = t/x$, 则

$$\theta = \int_0^1 x e^{-(xy)^2} dy$$

因此, $\theta = E_Y[xe^{-(xY)^2}]$, 其中$Y \sim U(0, 1)$. 从而产生$U(0, 1)$的$i.i.d$随机数$u_1, \cdots, u_m$, 则 θ的估计为

$$\hat{\theta} = \overline{g_m(u)} = \frac{1}{m} \sum_{i=1}^{m} x e^{-(xu)^2}.$$

此时对$x > 0$,$\Phi(x)$的估计为$0.5 + \hat{\theta}/\sqrt{2\pi}$; 对$x \leq 0$, 计算$\Phi(x) = 1 - \Phi(-x)$.

↑Example

```
x <- seq(.1, 2.5, length = 10)
m <- 10000
u <- runif(m)
```

```
cdf <- numeric(length(x))
for (i in 1:length(x)) {
    g <- x[i] * exp(-(u * x[i])^2 / 2)
    cdf[i] <- mean(g) / sqrt(2 * pi) + 0.5
}
Phi <- pnorm(x)
print(round(rbind(x, cdf, Phi), 3))
```

↓Example

例4: Monte Carlo积分, 无穷积分(续) 对上例, 我们可以使用另外一种方式(hit-or-miss). 记$Z \sim N(0,1)$, 则对任何实数x有

$$\Phi(x) = P(Z \le x) = EI(Z \le x).$$

从而从标准正态中产生随机样本$z_1, \cdots, z_m$后, 即可得到$\Phi(x)$的估计为

$$\hat{\Phi}(x) = \frac{1}{m} \sum_{i=1}^{m} I(z_i \le x).$$

其以概率1收敛到$\Phi(x)$.

↑Example

```
x <- seq(.1, 2.5, length = 10)
m <- 10000
z <- rnorm(m)
dim(x) <- length(x)
p <- apply(x, MARGIN = 1,
           FUN = function(x, z) {mean(z <= x)}, z = z)
Phi <- pnorm(x)
print(round(rbind(x, p, Phi), 3))
```

↓Example

总结如下: 欲估计积分

$$\theta = \int_A g(x)f(x)dx$$

其中f为以A为支撑的概率函数, 即$\int_A f(x)dx = 1$. 则产生f的$i.i.d$随机

数$x_1, \cdots, x_m$后, 由大数律知θ的估计为

$$\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} g(x_i).$$

由于$\hat{\theta}$的方差为σ^2/m, 其中$\sigma^2 = Var_f(g(X))$. 当随机变量X的分布未知时, 我们可以使用样本$x_1, \cdots, x_m$的经验分布函数, 从而得到σ^2/m的估计为

$$\hat{\sigma}^2/m = \frac{1}{m^2} \sum_{i=1}^{m} [g(x_i) - \overline{g(x)}]^2.$$

再由中心极限定理知道当$m \to \infty$时

$$\frac{\hat{\theta} - E\hat{\theta}}{\sqrt{Var(\hat{\theta})}}$$

依分布收敛到标准正态分布. 因此对大样本, 渐近正态性可以给出积分的Monte Carlo估计的误差界, 以此可以 来检查收敛性.

例5: Monte Carlo积分的误差界 建立$\Phi(2)$和$\Phi(2.5)$的Monte Carlo积分估计的%95置信区间.

↑Example

```
x <- 2
m <- 10000
z <- rnorm(m)
g <- (z <= x)  #the indicator function
v <- mean((g - mean(g))^2) / m
cdf <- mean(g)
c(cdf, v)
c(cdf - 1.96 * sqrt(v), cdf + 1.96 * sqrt(v))
#[1] 9.7800e-01 2.1516e-06
#[1] 0.975125 0.980875
```

↓Example

随机变量$I(Z \leq 2)$取值1的概率为$\Phi(2) \approx 0.977$. 此处$g(X) \sim B(10000, \Phi(2))$, 因此 $g(X)$的方差为$0.977(1 - 0.977)/10000 = 2.223e - 06$. Monte Carlo积分估计的方差为$2.1516e - 06$, 已经非常接近了.

1.1.2 Variance and Efficiency

Monte Carlo 方法在估计一个积分$\int_a^b g(x)dx$时, 将其表示为一个均匀随机变量的期望, 从而

$$\theta = \int_a^b g(x)dx = (b-a)\int_a^b g(x)\frac{1}{b-a}dx = (b-a)E[g(X)], \quad X \sim U(a,b).$$

从而算法如下

1. 从$U(a,b)$中产生$i.i.d$样本$X_1, \cdots, X_m$.
2. 计算$\overline{g(X)} = \frac{1}{m}\sum_{i=1}^m g(X_i)$.
3. $\hat{\theta} = (b-a)\overline{g(X)}$.

易知,

$$E\hat{\theta} = \theta, \quad Var(\hat{\theta}) = (b-a)^2 Var(\overline{g(X)}) = \frac{(b-a)^2}{m}Var(g(X).$$

有中心极限定理, $\overline{g(X)}$依分布渐近到正态分布, 因此$\hat{\theta}$也渐近到正态分布.

Hit-or-miss Monte Carlo 方法则使用了另外一种估计积分的方式, 其方差和上面说的方法不同. 表述如下: 假设f为随机变量X的概率函数, 使用"hit-or-miss"方法估计积分$F(x) = \int_{-\infty}^{x} f(t)dt$:

1. 从X的分布中产生$i.i.d$样本$X_1, \cdots, X_m$.
2. 计算$\overline{g(X)} = \frac{1}{m}\sum_{i=1}^{m} I(X_i \leq x)$.
3. $\hat{F(x)} = \overline{g(X)}$.

显然对每个有限的x, $Y = g(X) = I(X \leq x) \sim B(1, p), \quad p = F(x)$. 因此

$$E[\hat{F(x)}] = F(x), \quad Var[\hat{F(x)}] = F(x)(1 - F(x))/m.$$

$\hat{F(x)}$的方差可以通过$\hat{F(x)}(1 - \hat{F(x)})/m$来估计.

这两种方法的方差不同, 自然会问哪种优一些, 即更有效率.

设$\hat{\theta}_1$和$\hat{\theta}_2$是θ的两个无偏估计量，则$\hat{\theta}_1$比$\hat{\theta}_2$更有效，如果

$$Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2).$$

Definition

如果一个估计量的方差未知, 则我们通过样本将其估计出来. 另外, 估计量的方差总是可以通过增加样本量 来减少的.

1.2 Variance Reduction

我们已经知道Monte Carlo积分方法可以用来估计$E[g(X)]$这种类型的积分. 下面我们讨论如何减少$\theta = E[g(X)]$ 的样本均值估计量.

若$\hat{\theta}_1$和$\hat{\theta}_2$为参数θ的两个估计量，且$Var(\hat{\theta}_2) < Var(\hat{\theta}_1)$，则使用$\hat{\theta}_2$比使用$\hat{\theta}_1$方差的减少百分比为

$$100\left(\frac{Var(\hat{\theta}_1) - Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)}\right).$$

Definition

尽管可以增加样本量m来减少方差, 但是计算成本会很高. 比如标准差要至多为e, 而$Var(g(X)) = \sigma^2$, 则需要样本量$m \geq \frac{\sigma^2}{e^2}$. 如要把标准差从0.01减少到0.0001, 则样本量至少要10000.

1.3 Antithetic Variables

注意都对两个随机变量U_1, U_2,

$$Var((U_1 + U_2)/2) = \frac{1}{4}(Var(U_1) + Var(U_2) + 2Cov(U_1, U_2).$$

因此, 当U_1, U_2负相关时$(U_1 + U_2)/2$的方差比其相互独立时要小. 这个事实促使我们考虑使用负相关的随机变量 来减少方差.

比如, 假设$X_1, \ldots, X_n$为通过逆变换方法生产的, 对此m个随机向量$(X = (X_1, \ldots, X_n))$的每一个, 我们已经产生$U_j \sim U(0,1)$以及计算$X^{(j)} = F_X^{-1}(U_j), j = 1, \ldots, n$. 注意到若$U \sim U(0,1)$,则 $1 - U \sim U(0,1)$, 而U与$1 - U$负相关. 则

$$Y_j = g(X^{(j)}) = g(F_X^{-1}(U_1^{(j)}), \ldots, F_X^{-1}(U_n^{(j)}))$$

与

$$Y_j' = g(X^{(j)}) = g(F_X^{-1}(1 - U_1^{(j)}), \ldots, F_X^{-1}(1 - U_n^{(j)}))$$

同分布.

那么什么情形下Y_j与Y_j'是负相关的? 下面的定理说明如果g是单调的, 则Y_j和Y_j'是负相关的.

定义$(x_1, \ldots, x_n) \leq (y_1, \ldots, y_n)$ 如果$x_j \leq y_j, j = 1, \ldots, n$. 一个$n$元函数 $g = g(x_1, \ldots, x_n)$称为是增的, 如果其对每一个变量都是增的, 即如

果$(x_1,\ldots,x_n) \le (y_1,\ldots,y_n)$, 则$g(x_1,\ldots,x_n) \le g(y_1,\ldots,y_n)$. 类似的称$g$为递减的,如果其对每个变量都是减的. 最后, 称$g$是单调的,如果其是增的或者减的.

Theorem 1. 设$X = (X_1,\cdots,X_n)$各分量相互独立, f和g为同方向单调函数*(*同增或者同减*)*, 则

$$E[f(X)g(X)] \ge E[f(X)]E[g(X)].$$

Proof. 不妨设f,g同增. 证明方法是对n归纳. 假设$n = 1$, 则对所有的$x,y \in \mathcal{R}$, $(f(x)-f(y))(g(x)-g(y)) \ge 0$. 因此 $E[(f(X)-f(Y))(g(X)-g(Y))] \ge 0$, 即有

$$E[f(X)g(X)] + E[f(Y)g(Y)] \ge E[f(X)g(Y)] + E[f(Y)g(X)].$$

此处X,Y $i.i.d$, 因此

$$2E[f(X)g(X)] = E[f(X)g(X)] + E[f(Y)g(Y)]$$

$$\geq E[f(X)g(Y)] + E[f(Y)g(X)] = 2E[f(X)]E[g(X)],$$

即结论对$n = 1$成立. 现在假设对$n - 1$成立, 则由于

$$E[f(X)g(X)|X_n = x_n] \geq E[f(X_1, \ldots, X_{n-1}, x_n)]E[g(X_1, \ldots, X_{n-1}, x_n)]$$

$$= E[f(X|X_n = x_n)]E[g(X|X_n = x_n)],$$

最后一式每个为X_n的增函数, 因此根据$n = 1$的结论有

$$E[f(X)g(X)] = E\{E[f(X)g(X)|X_n]\}$$

$$\geq E\{E[f(X|X_n = x_n)]\}E\{E[g(X|X_n = x_n)]\}$$

$$= E[f(X)]E[g(X)].$$

□

Corollary 1. 设$g = g(X_1, \ldots, X_n)$ 是单调的, 则

$$Y = g(F_X^{-1}(U_1), \ldots, F_X^{-1}(U_n))$$

与

$$Y' = g(F_X^{-1}(1-U_1), \ldots, F_X^{-1}(1-U_n))$$

是负相关的.

Proof. 不失一般性, 假设g是递增的. 则

$$Y = g(F_X^{-1}(U_1), \ldots, F_X^{-1}(U_n))$$

与

$$-Y' = f = -g(F_X^{-1}(1-U_1), \ldots, F_X^{-1}(1-U_n))$$

是两个递增的函数. 因此$E[g(U)f(U)] \geq E[g(U)]E[f(U)]$, 即 $E[YY'] \leq E[Y]E[Y']$, 此即意味着

$$Cov(Y, Y') = E[YY'] - E[Y]E[Y'] \leq 0.$$

从而Y和Y'负相关. □

对偶变量(Antithetic Variables)方法很容易应用. 如果需要m个Monte Carlo重复, 则产生$m/2$个重复

$$Y_j = g(F_X^{-1}(U_1^{(j)}), \ldots, F_X^{-1}(U_n^{(j)}))$$

以及剩下$m/2$个重复

$$Y_j' = g(F_X^{-1}(1 - U_1^{(j)}), \ldots, F_X^{-1}(1 - U_n^{(j)})),$$

其中$U_i^{(j)}, i = 1, \ldots, n; j = 1, \ldots, m/2$. $i.i.d\ U(0,1)$. 则对偶估计量为

$$\begin{aligned} \hat{\theta} &= \frac{1}{m}\{Y_1 + Y_1' + Y_2 + Y_2' + \cdots + Y_{m/2} + Y_{m/2}'\} \\ &= \frac{2}{m}\sum_{j=1}^{m/2}\frac{Y_j + Y_j'}{2}. \end{aligned}$$

通过使用对偶变量使得得到的Monte Carlo 估计的方差减少.

例6: 对偶变量方法 使用之前的例子, 比如要估计积分

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{2\pi} e^{-t^2/2} dt.$$

使用对偶变量方法估计此积分, 并找出标准差的减少量.

作变量代换后, 目标积分是$\theta = E_U[xe^{-(xU)^2/2}]$, 其中$U \sim U(0,1)$. 显然此处$g$函数是单调的, 因此前面推论的条件满足. 产生随机数$u_1, \cdots, u_{m/2} \sim U(0,1)$, 计算

$$Y_j = g^{(j)}(u) = xe^{-(xu_j)^2/2}, j = 1, \cdots, m/2.$$

以及

$$Y_j' = xe^{-(x(1-u_j))^2/2}, j = 1, \cdots, m/2.$$

则样本均值

$$\hat{\theta} = \overline{g_m(u)} = \frac{1}{m/2} \sum_{j=1}^{m/2} \frac{xe^{-(u_j x)^2} + xe^{-(x(1-u_j))^2/2}}{2}$$

收敛到$E\hat{\theta}$.

从而若$x > 0$,则$\Phi(x)$的估计为$0.5+\hat{\theta}/\sqrt{2\pi}$; 若$x < 0$, 则$\Phi(x) = 1-\Phi(-x)$.

↑Example

```
MC.Phi <- function(x, R = 10000, antithetic = TRUE) {
u <- runif(R/2)
if (!antithetic) v <- runif(R/2) else
    v <- 1 - u
u <- c(u, v)
cdf <- numeric(length(x))
for (i in 1:length(x)) {
    g <- x[i] * exp(-(u * x[i])^2 / 2)
    cdf[i] <- mean(g) / sqrt(2 * pi) + 0.5
}
cdf
}
```

↓Example

和简单的Monte Carlo积分方法相比较:

↑Example

```
x <- seq(.1, 2.5, length=5)
Phi <- pnorm(x)
set.seed(123)
MC1 <- MC.Phi(x, anti = FALSE)
set.seed(123)
MC2 <- MC.Phi(x)
print(round(rbind(x, MC1, MC2, Phi), 5))
```

↓Example

方差的减少量可以模拟来比较:

↑Example

```
m <- 1000
MC1 <- MC2 <- numeric(m)
x <- 1.95
for (i in 1:m) {
    MC1[i] <- MC.Phi(x, R = 1000, anti = FALSE)
    MC2[i] <- MC.Phi(x, R = 1000)
}
print(sd(MC1))
print(sd(MC2))
print((var(MC1) - var(MC2))/var(MC1))
```

对$x = 1.95$, 对偶变量方法相比于简单Monte Carlo方法估计的方差大约减少99.5%.

1.4 Control Variates

Monte Carlo 估计中另外一种减少方差的方法是控制变量(Control Variates)的使用. 设要估计的量为$\theta = E[g(X)]$, f为一个函数, 其期望$E[f(X)] = \mu$已知, 且 f和g相关.

对任何常数c, 估计量$\hat{\theta}_c = g(X) + c(f(X) - \mu)$为无偏的, 且方差为

$$Var(\hat{\theta}_c) = Var(g(X)) + c^2 Var(f(X)) + 2cCov(g(X), f(X))$$

对c最小化, 则最小值在

$$c^* = -\frac{Cov(g(X), f(X))}{Var(f(X))}$$

处达到, 且最小值为

$$Var(\hat{\theta}_{c*}) = Var(g(X)) - \frac{[Cov(g(X), f(X))]^2}{Var(f(X))}.$$

随机变量$f(X)$称为$g(X)$的一个控制变量(Control Variate). 显然方差的减少率为

$$100\frac{[Cov(g(X), f(X))]^2}{Var(g(X))Var(f(X))} = 100[Cor(g(X), f(X))]^2.$$

可以看出, 这种方法在f和g强相关时是有优势的, 若f和g不相关, 则不会导致方差减少.

例7: 控制变量方法 使用控制变量方法计算积分

$$\theta = E[e^U] = \int_0^1 e^u du,$$

其中$U \sim U(0,1)$.

此例中,虽然积分值为$\theta = e - 1 \doteq 1.718282$, 我们仍然使用控制变量Monte Carlo方法来计算积分, 用以说明这种方法的使用. 如果使用简单的Monte

Carlo积分方法, 则方差为

$$Var(g(U)) = Var(e^U) = \frac{e^2-1}{2} - (e-1)^2 \doteq 0.2420351.$$

重复m次得到的估计的方差为$Var(g(U))/m$.

控制变量的自然选择为$U \sim U(0,1)$, 则$Cov(e^U, U) = 1-(e-1)/2 \doteq 0.1408591$. 因此

$$c^* = \frac{-Cov(e^U, U)}{Var(U)} = -12 + 6(e-1) \doteq -1.690309.$$

而使用控制变量方法得到的估计为$\hat{\theta}_{c^*} = e^U - 1.690309(U-0.5)$, m次重复后的方差$Var(\hat{\theta}_{c^*})/m$, 其中$Var(\hat{\theta}_{c^*})$为

$$Var(e^U) - \frac{[Cov(e^U,U)]^2}{Var(U)} = \frac{e^2-1}{2} - (e-1)^2 - 12[1-\frac{e-1}{2}]^2 \doteq 0.003940175.$$

因此使用控制变量方法导致简单Monte Carlo估计量的方差减少率为$100(1 - 0.003940175/0.2429355) = 98.3781\%$.

下面我们使用控制变量方法来计算其经验的方差减少率.

↑Example

```
x <- seq(.1, 2.5, length=5)
Phi <- pnorm(x)
set.seed(123)
MC1 <- MC.Phi(x, anti = FALSE)
set.seed(123)
MC2 <- MC.Phi(x)
print(round(rbind(x, MC1, MC2, Phi), 5))
```

↓Example

方差的减少量可以模拟来比较:

↑Example

```
m <- 10000
a <- - 12 + 6 * (exp(1) - 1)
U <- runif(m)
T1 <- exp(U)                    #simple MC
T2 <- exp(U) + a * (U - 1/2)  #controlled
mean(T1)
```

```
mean(T2)
(var(T1) - var(T2)) / var(T1)
```

↓Example

例8: 使用控制变量方法的Monte Carlo积分 使用控制变量方法计算积分

$$\int_0^1 \frac{e^{-x}}{1+x^2}dx.$$

此例中感兴趣的量为$\theta = Eg(X), g(x) = \frac{e^{-x}}{1+x^2}$, 其中$X \sim U(0,1)$. 我们 要寻求一个足够接近$g$的函数$f$且其期望值要已知, 和$g$相关. 比如$f(x) = \frac{e^{-0.5}}{(1+x^2)}$ 是可以的, 若$U \sim U(0,1)$, 则

$$Ef(U) = e^{-0.5}\int_0^1 \frac{1}{1+u^2}du = e^{-0.5}\frac{\pi}{4}.$$

我们也可以估计出$Cor(g(U), f(U)) \approx 0.974$. 因此

↑Example

```
f <- function(u)  exp(-.5)/(1+u^2)
g <- function(u)  exp(-u)/(1+u^2)
set.seed(510) #needed later
u <- runif(10000)
B <- f(u)
A <- g(u)
a <- -cov(A,B) / var(B)    #est of c*

m <- 100000
u <- runif(m)
T1 <- g(u)
T2 <- T1 + a * (f(u) - exp(-.5)*pi/4)
c(mean(T1), mean(T2))
c(var(T1), var(T2))
(var(T1) - var(T2)) / var(T1)
```

↓Example

1.4.1 Antithetic variate as control variate

对偶变量方法实际上是控制变量方法的特例. 注意到控制变量方法 是无偏估计的线性组合. 一般地, 若$\hat{\theta}_1$和$\hat{\theta}_2$ 为θ的无偏估计量, 则对任何常数c,有

$$\hat{\theta}_c = c\hat{\theta}_1 + (1-c)\hat{\theta}_2$$

仍为θ的无偏估计. 其方差为

$$Var(\hat{\theta}_2) + c^2 Var(\hat{\theta}_1 - \hat{\theta}_2) + 2cCov(\hat{\theta}_2, \hat{\theta}_1 - \hat{\theta}_2).$$

特别地, 当$\hat{\theta}_1$和$\hat{\theta}_2$同分布, 且$Cor(\hat{\theta}_1, \hat{\theta}_2) = -1$. 则$Cov(\hat{\theta}_1, \hat{\theta}_2) = -Var(\hat{\theta}_1)$, 方差为

$$Var(\hat{\theta}_c) = 4c^2 Var(\hat{\theta}_1) - 4cVar(\hat{\theta}_1) + Var(\hat{\theta}_1) = (4c^2 - 4c + 1)Var(\hat{\theta}_1),$$

因此最优的$c^* = 1/2$. 此时控制变量估计量为

$$\hat{\theta}_{c^*} = \frac{\hat{\theta}_1 + \hat{\theta}_2}{2}$$

这也是(这种特定选择下)对偶变量方法下的估计量.

1.4.2 Several control variates

将无偏估计量组合起来作为参数θ的估计,以减少方差的方法可以推广到多个控制变量场合:

$$\hat{\theta}_c = g(X) + \sum_{i=1}^{k} c_i(f_i(X) - \mu_i))$$

其中$\mu_i = Ef_i(X), i = 1, \cdots, k.$, $\sum_{i=1}^{k} c_i = 1$ 以及

$$E\hat{\theta}_c = E[g(X)] + \sum_{i=1}^{k} c_i E[f_i(X) - \mu_i] = \theta$$

控制变量方法下的估计量$\hat{\theta}_{\hat{c}}$以及最优的c_i可以通过回归模型来估计.

在 $k = 1$ 场合, 考虑二元样本$((g(X_1), f(X_1)), \cdots, (g(X_n), f(X_n)))$, 假设$g(X)$与$f(X)$之间存在线性关系: $g(X) = \beta_0 + \beta_1 f(X) + e$, 且

$$E[g(X)] = \beta_0 + \beta_1 E[f(X)].$$

β_1的最小二乘估计为

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (g(X_i) - g(\bar{X}))(f(X_i) - f(\bar{X}))}{\sum_{i=1}^n (f(X_i) - f(\bar{X}))^2} = \frac{\hat{Cov}(g(X), f(X))}{\hat{Var}(f(X))} = -c^*.$$

这说明我们可以使用$g(X)$对$f(X)$的回归来估计c:

↑Code

```
L<-lm(gx~fx)
c.star<--L$coef[2]
```

↓Code

截距的最小二乘估计为$\hat{\beta}_0 = \overline{g(X)} - (-c^*)\overline{f(X)}$, 因此在$\mu$处的预测值为

$$\hat{\beta}_0 + \hat{\beta}_1 \mu = \overline{g(X)} + \hat{c}^*(\overline{f(X)} - \mu) = \hat{\theta}_{c^*}$$

即控制变量方法下的估计量是预测值.

误差方差的估计为

$$\hat{\sigma}_e^2 = \hat{Var}(g(X) - \hat{g}(X)) = \hat{Var}(g(X) - (\beta_0 + \beta_1 f(X))) = \hat{Var}(g(X) + c^* f(X))$$

控制变量方法下的估计量的方差估计为

$$\hat{Var}(\overline{g(X)} + \hat{c}^*(\overline{f(X)} - \mu)) = \frac{g(X) + \hat{c}^* f(X)}{n} = \frac{\hat{\sigma}_e^2}{n}.$$

因此在R中, 控制变量方法下的估计量的标准差的估计为

↑Code

```
se.hat<-summary(L)$sigma/sqrt(n)
```

↓Code

和前面控制变量一节中的结果相同.

对一般的k, 则可以使用回归

$$g(X) = \beta_0 + \sum_{i=1}^{k} \beta_i f(X) + e$$

来估计最优的常数$c^* = (c_1^*, \cdots, c_k^*)$. 则$-c^* = (\hat{\beta}_1, \cdots, \hat{\beta}_k)$, 以及此时控制变量方法下的估计为在$\mu = (\mu_1, \cdots, \mu_k)$处的预测值$\hat{g}(X)$. 估计的方差为$\sigma_e^2/n$.

例9: 控制变量和回归 使用回归方法估计积分

$$g(x) = \int_0^1 \frac{e^{-x}}{1+x^2} dx.$$

这里控制变量取$f(x) = e^{.5}(1+x^2)^{-1}, 0 < x < 1.$, $\mu = E[f(X)] = e^{.5}\pi/4.$ 为估计最优常数c^*,

↑Example

```
set.seed(510)
u <- runif(10000)
f <- exp(-.5)/(1+u^2)
g <- exp(-u)/(1+u^2)
L <- lm(g~f)
c.star <-  - L$coeff[2]   # beta[1]
mu <- exp(-.5)*pi/4
c.star
theta.hat <- sum(L$coeff * c(1, mu))  #pred. value at mu
theta.hat
```

```
summary(L)$sigma^2
summary(L)$r.squared
```

↓Example

这里我们使用了和前例中同样的种子, 因此得到同样的c^*估计. 现在$\hat{\theta}_{\hat{c}^*}$是在$\mu = 0.4763681$处的预测值. 而估计量$\hat{\theta}$及其标准差, 方差的减少率都和前例相同.

1.5 Importance sampling

在前面内容中,我们将有限区间上的积分视为是此区间上的均匀分布随机变量的某个函数的期望值. 这种方法的缺点是不能直接用于无穷积分的估计, 以及当被积函数在积分区间上不是很均匀的时候 效率很低. 那么很自然地可以考虑均匀分布以外的其他加权函数. 这就导致“重要性抽样方法”.

假设随机变量X的密度函数为$f(x)$, 满足在集合$\{x : g(X) > 0\}$上$f(x) > 0$. 则记$Y = g(X)/f(X)$,

$$\int g(x)dx = \int \frac{g(x)}{f(x)} f(x)dx = EY.$$

通过简单的Monte Carlo方法估计EY:

$$\frac{1}{m}\sum_{i=1}^{m} Y_i = \frac{1}{m}\sum_{i=1}^{m} \frac{g(X_i)}{f(X_i)}.$$

此处随机变量$X_1, \cdots, X_m$为从f中抽取的样本. f称为**重要函数**. 估计的 方差为$Var(Y)/m$, 当Y接近常数时, Y的方差会比较小, 此时密度f应该"靠近" 函数$g(x)$. 当然, 密度f应该选择容易抽样的分布.

重要性抽样方法的优点是可以通过选择重要函数来降低 Monte Carlo 估计的方差. 假设$f(x)$是支撑为A的密度函数, 若$\phi(x) > 0$, $\forall x \in A$, 则积分

$$\theta = \int_A g(x)f(x)dx,$$

可以被写为

$$\theta = \int_A g(x)\frac{f(x)}{\phi(x)}\phi(x)dx.$$

若$\phi(x)$为A上的密度函数, 则θ的一个估计量为

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)\frac{f(X_i)}{\phi(X_i)},$$

其中$X_1, \cdots, X_n$为从$\phi(x)$中产生的随机样本. $\phi(x)$称为重要性抽样函数或者 包络函数(envelope). ϕ的选择有各种各样的, 典型的选择使得在A上有$\phi(x) \approx |g(x)|f(x)$, 当然$\phi$有有限的方差. 这是由于

$$Var(\hat{\theta}) = \int_A \frac{g^2(x)}{\phi(x)} dx - \theta^2.$$

由Cauchy-Schwarz不等式, $Var(\hat{\theta})$的最小值为

$$\left(\int_A |g(x)| dx \right)^2 - \theta^2.$$

可以通过取

$$\phi(x) = \frac{|g(x)|}{\int_A |g(x)| dx}.$$

不过由于要估计$\int_A |g(x)|dx$, 因此分母未知. 尽管这样能达到最小方差的$\phi(x)$很难取到, 但是如果选择和$|g(x)|$形状接近的$\phi(x)$, 则方差可以接近最优方差.

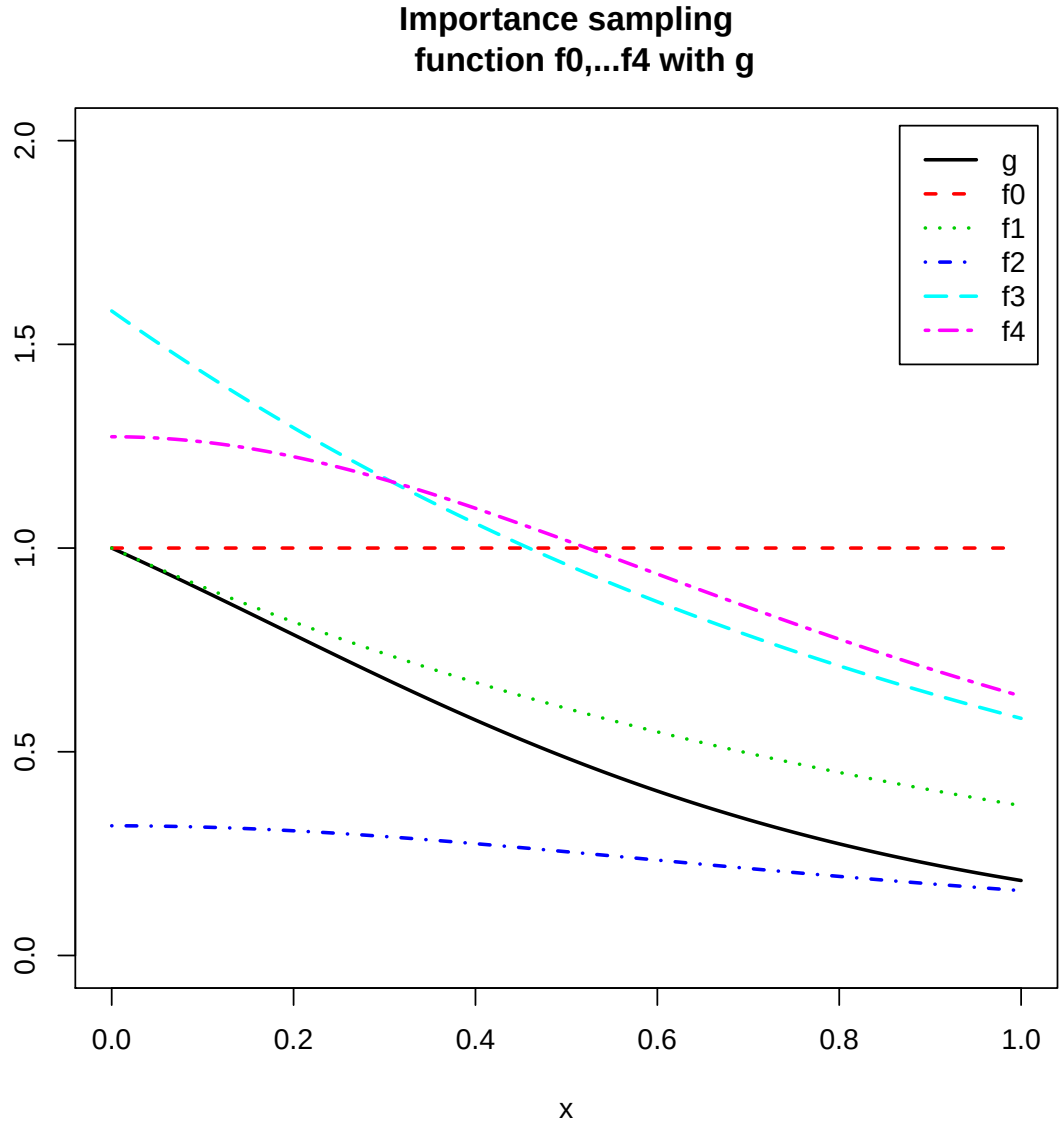
例10: 重要性函数的选择 使用不同的重要性函数来估计积分

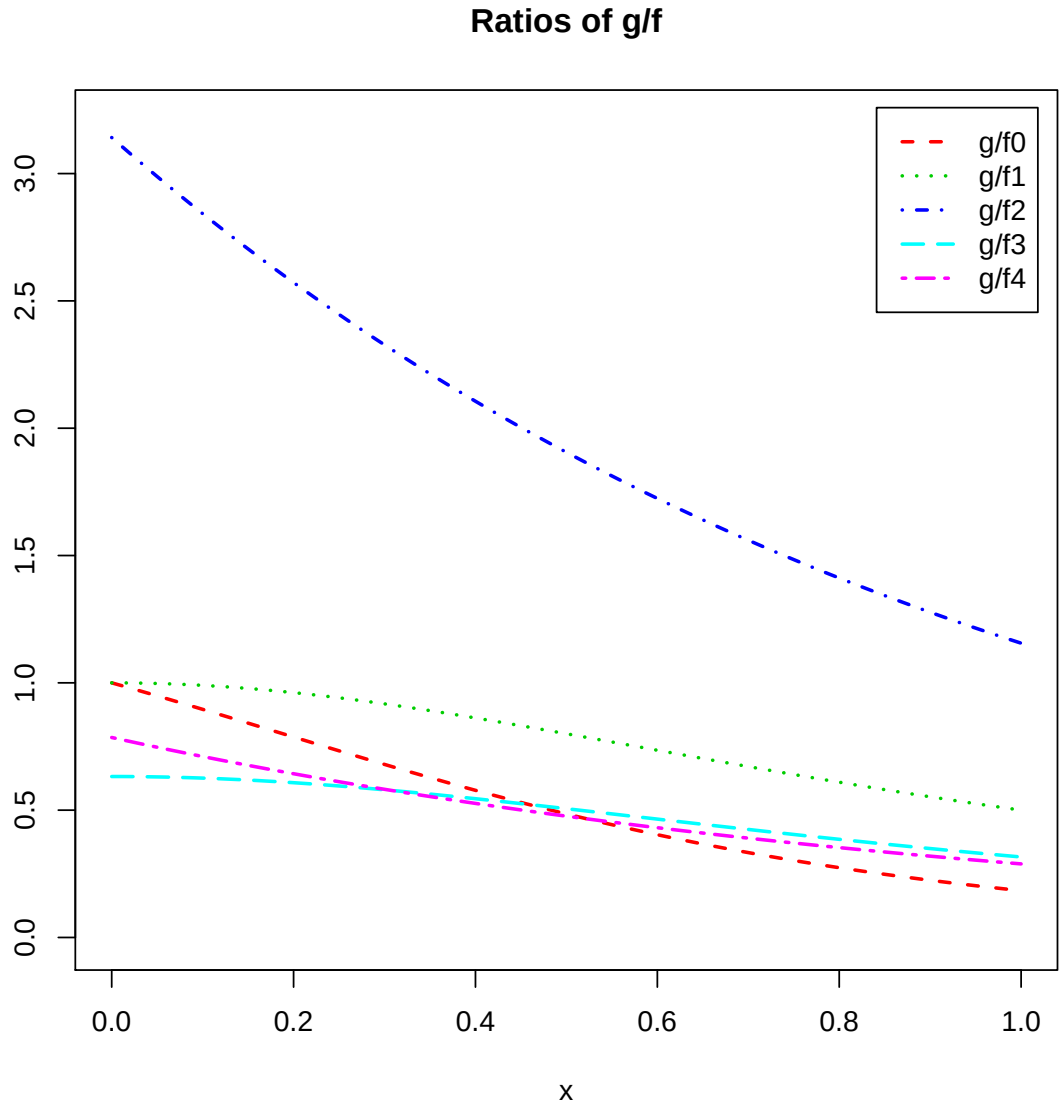
$$\int_0^1 \frac{e^{-x}}{1+x^2}dx.$$

比如, 可以考虑如下重要性函数:

$$\begin{aligned}
f_0(x) &= 1,\ 0<x<1,\\
f_1(x) &= e^{-x},\ 0<x<\infty,\\
f_2(x) &= \frac{1}{\pi(1+x^2)},\ -\infty<x<\infty,\\
f_3(x) &= e^{-x}(1-e^{-1})^{-1},\ 0<x<1,\\
f_4(x) &= \frac{4}{\pi(1+x^2)},\ 0<x<1
\end{aligned}$$

这5种函数和g之间的关系:


Importance sampling
function f0,...f4 with g
g
f0
f1
f2
f3
f4
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
x


Ratios of g/f
g/f0
g/f1
g/f2
g/f3
g/f4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0
x

在这5种重要性函数中, f_1, f_2的支撑比$(0, 1)$大, 因此很多样本都对估计积分无用, 因此效率低下. f_3最接近g. 比较这5种重要性函数下的积分估计值和方差估计值:

↑Example

```
m <- 10000
theta.hat <- se <- numeric(5)
g <- function(x) {
    exp(-x - log(1+x^2)) * (x > 0) * (x < 1)
    }

x <- runif(m)      #using f0
fg <- g(x)
theta.hat[1] <- mean(fg)
se[1] <- sd(fg)

x <- rexp(m, 1)    #using f1
fg <- g(x) / exp(-x)
theta.hat[2] <- mean(fg)
se[2] <- sd(fg)
```

```
x <- rcauchy(m)   #using f2
i <- c(which(x > 1), which(x < 0))
x[i] <- 2  #to catch overflow errors in g(x)
fg <- g(x) / dcauchy(x)
theta.hat[3] <- mean(fg)
se[3] <- sd(fg)

u <- runif(m)     #f3, inverse transform method
x <- - log(1 - u * (1 - exp(-1)))
fg <- g(x) / (exp(-x) / (1 - exp(-1)))
theta.hat[4] <- mean(fg)
se[4] <- sd(fg)

u <- runif(m)    #f4, inverse transform method
x <- tan(pi * u / 4)
fg <- g(x) / (4 / ((1 + x^2) * pi))
theta.hat[5] <- mean(fg)
se[5] <- sd(fg)

rbind(theta.hat, se)
```

```
                    [,1]      [,2]      [,3]       [,4]      [,5]
theta.hat 0.5243630 0.5188012 0.5367005 0.52389845 0.5259339
se        0.2436977 0.4199031 0.9594029 0.09657142 0.1414429
>
```

↓Example

1.6 Stratified Sampling

利用积分的线性性和和强大数律, 可以知道要求的积分$\int g(x)dx$可以分为几个积分之和, 对每个积分可以单独使用Monte Carlo积分方法. 不妨设将此积分分为k个积分之和, 对每个积分 抽样m_i次, $m = m_1 + \cdots + m_k$, 满足目标

$$Var(\hat{\theta}_k(m_1, \cdots, m_k)) < Var(\hat{\theta}).$$

例11: 考虑之前的例子 将积分区间$(0, 1)$比如分为四个子区间, 在每个子区间上抽样$m/4$次, 这里m为总的抽样次数. 然后将这些估计加起来得到积分$\int_0^1 e^{-x}(1 + x^2)^{-1}dx$的估计.

↑Example

```
M <- 20   #number of replicates
T2 <- numeric(4)
estimates <- matrix(0, 10, 2)

g <- function(x) {
    exp(-x - log(1+x^2)) * (x > 0) * (x < 1) }

for (i in 1:10) {
    estimates[i, 1] <- mean(g(runif(M)))
    T2[1] <- mean(g(runif(M/4, 0, .25)))
    T2[2] <- mean(g(runif(M/4, .25, .5)))
    T2[3] <- mean(g(runif(M/4, .5, .75)))
    T2[4] <- mean(g(runif(M/4, .75, 1)))
    estimates[i, 2] <- mean(T2)
}

estimates
apply(estimates, 2, mean)
apply(estimates, 2, var)
```

↓Example

Theorem 2. 记M次抽样下的简单*Monte Carlo*积分估计量*(*即从均匀分布中抽样*)*为$\hat{\theta}^M$, 以及

$$\hat{\theta}^S = \frac{1}{k}\sum_{i=1}^{k}\hat{\theta}_i$$

表示分层的估计量, 每个层的抽样次数为m/k. 在第j个层上, $g(U)$的均值和方差分别 记为θ_j和σ_j^2, $j = 1, \cdots, k$. 则$Var(\hat{\theta}^M) \geq Var(\hat{\theta}^S)$.

Proof. 用J表示随机选择的层, $P(J = j) = 1/k, j = 1, \cdots, k$. 则

$$\begin{aligned}
Var(\hat{\theta}^M) &= \frac{Var(g(U))}{M} = \frac{1}{M}[Var(E(g(U|J))) + E(Var(g(U|J)))] \\
&= \frac{1}{M}(Var(\theta_J) + E\sigma_J^2) \\
&= \frac{1}{M}(Var(\theta_J) + \frac{1}{k}\sum_{i=1}^{k}\sigma_j^2) \\
&= \frac{1}{M}Var(\theta_J) + Var(\hat{\theta}^S) \geq Var(\hat{\theta}^S).
\end{aligned}$$

□

例12: 使用分层抽样方法估计前面例子中的积分 对积分$\int_0^1 e^{-x}(1+x^2)^{-1}dx$应用 k层抽样方法估计, 并和简单的Monte Carlo积分方法比较.

↑Example

```
M <- 10000  #number of replicates
   k <- 10     #number of strata
   r <- M / k  #replicates per stratum
   N <- 50     #number of times to repeat the estimation
   T2 <- numeric(k)
   estimates <- matrix(0, N, 2)

   g <- function(x) {
       exp(-x - log(1+x^2)) * (x > 0) * (x < 1)
       }

   for (i in 1:N) {
       estimates[i, 1] <- mean(g(runif(M)))
       for (j in 1:k)
           T2[j] <- mean(g(runif(M/k, (j-1)/k, j/k)))
       estimates[i, 2] <- mean(T2)
   }
```

```
apply(estimates, 2, mean)
apply(estimates, 2, var)
```

↓Example

1.7 Stratified Importance Sampling

分层抽样的思想可以用在重要性抽样方法中. 在重要性抽样方法中, $\theta = \int g(x)dx$的估计量方差为σ^2/M, 其中$\sigma^2 = Var(g(X)/f(X)), X \sim f$. 应用分层抽样方法, 将直线分为$k$个子区间 $I_j = \{x : a_{j-1} \le x < a_j\}$, 其中 $a_0 = -\infty, a_j = F^{-1}(j/k), j = 1, \cdots, k-1, a_k = \infty$. 记$g_j(x) = g(x)I(a_{j-1} \le x < a_j)$, 以及

$$\theta_j = \int_{a_{j-1}}^{a_j} g_j(x)dx, j = 1, \cdots, k.$$

则$\theta = \theta_1 + \cdots + \theta_k$. 对每个子区间$I_j$, 重要性函数可以取为条件密度:

$$f_j(x) = f(x|I_j) = \frac{f(x)I(a_{j-1} \le x < a_j)}{P(a_{j-1} \le X < a_j)} = kf(x)I(a_{j-1} \le x < a_j)$$

再记$\sigma_j^2 = Var(g_j(X)/f_j(X)), j = 1, \cdots, k$. 然后得到$\theta$的分层重要性抽样下的估计为

$$\hat{\theta}^{SI} = \sum_{i=1}^{k} \hat{\theta}_j$$

其方差为

$$Var(\hat{\theta}^{SI}) = \sum_{i=1}^{k} Var(\hat{\theta}_j) = \frac{1}{m} \sum_{i=1}^{k} \sigma_j^2.$$

其中$m = M/k$. 则希望

$$\sigma^2/M > \frac{k}{M} \sum_{i=1}^{k} \sigma_j^2.$$

我们可以证明如下结论:

Theorem 3. 假设$M = km$为重要性抽样下的估计量$\hat{\theta}^I$的抽样个数, $\hat{\theta}^{SI} = \sum_{i=1}^{k} \hat{\theta}_j$为分层重要性抽样下的估计量, 这里$\hat{\theta}_j$为第$j$层$\theta_j$的重要性抽样下的估计量, 抽样个数为$m$. 若$Var(\hat{\theta}^I) = \sigma^2/M$, 以及$Var(\hat{\theta}_j) = \sigma_j^2/m$,则

$$\sigma^2 - k\sum_{j=1}^{k} \sigma_j^2 \geq 0,$$

等号成立当且仅当$\theta_1 = \cdots = \theta_k$.

此结论说明分层抽样绝不会扩大方差, 在g非常数的场合总是存在一个可以减少方差的分层.

例13: 在前面的例子中, 试使用$f_3(x)$作为重要性函数, 将积分区间分为$(j/5, (j+1)/5), j = 0, 1, \cdots, 4$这 5个子区间, 对每个子区间应用重要性抽样方法, 计算此时积分的估计量及其经验方差.