STAT 3006: Statistical Computing
Lecture 6*

12 February

Described in the previous lecture, the accept-reject method proceeds as follows.

- Generate $Y \sim g$ and $U \sim unif[0,1]$,
- Accept $X = Y$ if $U \leq \frac{f(Y)}{Mg(Y)}$.

The pdf g should be easily sampled, close to f, and has a heavier tail and a larger domain than those of f. Notice that the acceptance rate is $\frac{1}{M}$, so we want to make M as small as possible. Ideally, we can let M be $\max_y \frac{f(y)}{g(y)}$. If $\max_y \frac{f(y)}{g(y)} = \infty$, we have to give up the g and choose another easily sampled pdf. For example, when f is the pdf of t distribution with two degrees of freedom and g is the pdf of the standard normal, $\lim_{y \to +\infty} \frac{f(y)}{g(y)} = \infty$ and $\lim_{y \leftarrow -\infty} \frac{f(y)}{g(y)} = \infty$. In this case, we cannot use the standard normal as a proposal distribution for sampling t_2 distribution because of very tight tails of the standard normal. Actually, the Cauchy distribution can be used to sample from t_2.

5 Parameter Estimation Techniques

We have known how to sample from f in the last lecture, in this lecture, we focus on estimating a parameter associated with f, e.g., $E_f(h(X)) = \int h(x)f(x)dx$ where $h(x)$ is given. For illustration,

- $h(x) = x$, $E_f(h(X)) = \mu$, the mean of X;
- $h(x) = x^2$, $E_f(h(X)) = \sigma^2 + \mu^2$, the summation of squared mean and variance of X;
- $h(x) = I(X \geq 50)$, $E_f(h(X)) = P(X \geq 50)$, the probability that X is greater than 50.

Sometimes numerically calculating the integral $\int h(x)f(x)dx$ is difficult and leads to unstable results. In statistics, there are some approaches to overcome the problem.

*If you have any question about the note, please send an email to xyluo@link.cuhk.edu.hk

5.1 Classic Monte Carlo Integration

Draw samples $X_1, \ldots, X_n$ i.i.d. $\sim f(x)$, then use $\frac{\sum_{i=1}^n h(X_i)}{n}$ to approximate $E_f(h(X))$. According to strong law of large numbers, $\lim_{n\to\infty} \frac{\sum_{i=1}^n h(X_i)}{n} = E_f(h(X))$ almost surely. As long as we collect enough samples from $f(x)$, we can have an accurate estimate for $E_f(h(X))$.

What if it is hard to sample from $f(x)$? We must ask the inverse method or the accept-reject method for help.

5.2 Importance Sampling

Definition 5.1. Given a proposal distribution g, we collect n samples $x_1, \ldots, x_n$ from g and approximate $E_f(h(X))$ by $\frac{1}{n}\sum_{i=1}^n \frac{f(x_i)}{g(x_i)} h(x_i)$.

Remarks:

- $E_f(h(X)) = \int h(x)f(x)dx = \int \frac{f(x)}{g(x)} h(x)g(x)dx = E_g(\frac{f(X)}{g(X)} h(X))$, so we can use samples from g to approximate $E_f(h(X))$.
- the range of g should be larger than that of f.
- sample once (m=100), we can estimate multiple $E_f(h_1(X))$, $E_f(h_2(X))$.
- sample once (m=100), we can estimate multiple $E_{f_1}(h(X))$, $E_{f_2}(h(X))$.

Example: $X \sim Binomial(m,p)$ where $m = 10,000$, $p = 1\%$ (rare disease), we want to calculate $P(X \geq 500)$. Notice that $500 >> mp = 100$. Approach 1: independently draw samples $\sim Binomial(m,p)$, then count $\frac{1}{n}\sum_{i=1}^n I(X_i \geq 500)$, which is not efficient. Approach 2: (Importance Sampling) $h(x) = I(x \geq 500)$, $f(x) = \binom{m}{x} p^x (1-p)^{m-x}$, $g(x) = \binom{m}{x} q^x (1-q)^{m-x} (q = 0.9)$, draw samples from $g(x)$, then count $\frac{1}{n}\sum_{i=1}^n I(X_i \leq 500)(\frac{p}{q})^{X_i}(\frac{1-p}{1-q})^{m-X_i}$.

How to choose $g(x)$ in the importance sampling. Let $m(X_1, \ldots, X_n)$ be $\frac{1}{n}\sum_{i=1}^n \frac{f(X_i)}{g(X_i)} h(X_i)$ where $X_i \sim g(x)$.

$$\begin{aligned} Var(m(X_1, \ldots, X_n)) &= \frac{1}{n}\left[E_g(h^2 \frac{f^2}{g^2}) - (E_g(h\frac{f}{g}))^2\right] \\ &= \frac{1}{n}\left[E_g(h^2 \frac{f^2}{g^2}) - E_f(h)^2\right]. \end{aligned} \tag{5.1}$$

Therefore, $Var(m(X_1, \ldots, X_n)) < \infty$ implies $E_g(h^2 \frac{f^2}{g^2})\infty$. Notice that $E_g(h^2 \frac{f^2}{g^2}) = \int h^2(x)f(x)\frac{f(x)}{g(x)}dx$.

Sufficient condition for $Var(m(X_1, \ldots, X_n)) < \infty$:

1 $E_g(h^2(X)) < \infty$.

2 $\frac{f(x)}{g(x)} < M$, $\forall x$.

Theorem 5.1. $g^*(x) = \frac{|h(x)|f(x)}{\int |h(t)|f(t)dt}$ *minimizes the* $Var(m(X_1, \ldots, X_n))$.

Proof. Due to Equation (5.1), minimizing $Var(m(X_1, \ldots, X_n))$ (w.r.t g) is equivalent to minimizing $E_g(h^2 \frac{f^2}{g^2})$ (w.r.t. g). Jensen's inequality implies that $E_g(h^2 \frac{f^2}{g^2}) \geq (E_g(\frac{|h|f}{g}))^2 = (\int |h| f dx)^2$. The equality holds if and only if $\frac{|h(x)|f(x)}{g^*(x)} = \int |h(x)|f(x)dx$. It follows that the $g^*(x) = \frac{|h(x)|f(x)}{\int |h(t)|f(t)dt}$ minimizes $Var(m(X_1, \ldots, X_n))$.

5.3 Stratified Sampling

The whole population can be partitioned into strata $S_1, S_2, \ldots, S_m$. Sample $X_{i1}, X_{i2}, \ldots, X_{in_i}$ from the stratum S_i. $E(h(X)|X \in S_i) \approx \frac{1}{n_i} \sum_{j=1}^{n_i} h(X_{ij})$. $E_f(h(X)) \approx \sum_{i=1}^{m} \mu(S_i) \frac{1}{n_i} \sum_{j=1}^{n_i} h(X_{ij}) := m(\mathbf{X})$.

$$
\begin{aligned}
Var(m(\mathbf{X})) &= \sum_{i=1}^{m} \mu(S_i)^2 Var(\frac{1}{n_i} \sum_{j=1}^{n_i} h(X_{ij})) \\
&= \sum_{i=1}^{m} \frac{\mu^2(S_i)}{n_i} Var(h(X)|X \in S_i).
\end{aligned}
$$

Question: How to choose $n_1, \ldots, n_m$ and $n_1 + \ldots + n_m = n$ to minimize $Var(m(\mathbf{X}))$?
Denote $Var(h(X)|X \in S_i)$ by V_i. The problem is minimizing $\sum_{i=1}^{m} \frac{\mu^2(S_i)V_i}{n_i}$ subject to $n_1 + n_2 + \ldots + n_m = n$. Let $L = \sum_{i=1}^{m} \frac{\mu^2(S_i)V_i}{n_i} - \lambda(n - n_1 - \ldots - n_m)$. Based on Lagrange multiplies,

$$
\begin{aligned}
\frac{\partial L}{\partial n_i} &= -\frac{\mu^2(S_i)V_i}{n_i^2} + \lambda = 0 \\
n_i &= \sqrt{\frac{\mu^2(S_i)V_i}{\lambda}} \\
\frac{\sum_{i=1}^{m} \sqrt{\mu^2(S_i)V_i}}{\sqrt{\lambda}} &= n \\
\sqrt{\lambda} &= \frac{\sum_{i=1}^{m} \mu(S_i)\sqrt{V_i}}{n} \\
n_i &= \frac{n\mu(S_i)\sqrt{V_i}}{\sum_{i=1}^{m} \mu(S_i)\sqrt{V_i}}. \qquad (5.2)
\end{aligned}
$$