Computational Statistics
4. exercise session, 19.2.2010

1. (A "Rao-Blackwellized" density estimate.) Suppose we generate a sample $(X_i, Y_i), i = 1, \ldots, N$ for a two-dimensional continuous distribution. Suppose also that the conditional density $f_{X|Y}$ is available. Now we explore two approaces for estimating the marginal density f_X of variable X. The first approach is to apply a nonparametric density estimation method (such as a density histogram or a kernel estimate) to the values (X_i), which (of course) come from the marginal density f_X. The second approach is to notice that

$$f_X(x) = \int f_{X,Y}(x, y)\,\mathrm{d}y = \int f_{X|Y}(x \mid y)\, f_Y(y)\,\mathrm{d}y \approx \frac{1}{N}\sum_{i=1}^{N} f_{X|Y}(x \mid Y_i).$$

The last approximation is ordinary Monte Carlo integration, since (Y_i) is a sample from the marginal density f_Y. Some call this approach "Rao-Blackwellization" since it is similar in spirit to conditioning (but not literally the same).

Begin by simulating $n = 100$ values using the following code

```
n <- 100
k <- 4
y <- rgamma(n, k/2, k/2)
x <- rnorm(n, 0, sd = 1 / sqrt(y))
```

Pretend that you do not notice that the marginal distribution of X is the t distribution with $k = 4$ degrees of freedom.

Now, estimate f_X by plotting a probability density histogram and a kernel density estimate of the simulated x values. In R, the kernel density estimate can be drawn with

```
plot(density(x), xlim = c(-10, 10))
```

You can also try to change the bandwidth of the kernel estimate manually, e.g., `plot(density(x, bw = 0.5), xlim = c(-10, 10))`. Secondly, draw a Rao–Blackwellized density estimate of $f_X(x)$ on the interval $(-10, 10)$. (You need to set up a grid on that interval and calculate the average of normal density values evaluated on the grid.)

2. Let $Z \sim N(0, 1)$. Estimate $I = E|Z|^{2.3}$ firstly, by naive Monte Carlo integration, and secondly by using Z^2 as the control variate (obviously, $EZ^2 = 1$).

Calculate also Monte Carlo standard errors for the two methods. Use a moderate-sized Monte Carlo sample (e.g., $N = 10000$) and a separate, smaller pilot sample for the control variate method. (The expectation I can be expressed using the gamma function, but you are not expected to do that in this problem.)

3. (Using importance sampling for rare event simulation.) Estimate

$$P(Z \geq 5), \qquad \text{where} \quad Z \sim N(0,1)$$

firstly by naive Monte Carlo, where you simulate from $N(0,1)$, and secondly by importance sampling, where you generate the draws with

$$X_i = 5 + Y_i, \qquad \text{where} \quad Y_i \sim \text{Exp}(5).$$

(This choice is inspired by our previous analysis of the truncated normal distribution.) Use $N = 10000$. Estimate the Monte Carlo standard error for both methods.

4. (A good and a bad instrumental density in importance sampling). Suppose that $X \sim N(0, 10^2)$ and that we want to estimate $I = EX^2$ $(= 100)$ using importance sampling.

Calculate the estimate of I and also its (Monte Carlo) standard error, when the instrumental distribution is

- the Cauchy distribution,
- the standard normal distribution $N(0,1)$.

Use a moderate sample size such as $N = 10000$.

To understand concretely why the first approach succeeds and the second approach fails miserably, check what is the maximum value of $|X_i|$ that you obtained in your two samples.

5. To understand theoretically what went wrong while trying to use $N(0,1)$ as the instrumental distribution in the previous problem, find for which values of $\sigma > 0$ the integral

$$\int_{-\infty}^{\infty} x^4 \, \frac{f(x)^2}{g(x \mid \sigma)} \, \mathrm{d}x$$

is finite, when $f(x) = N(x \mid 0, 10^2)$ and $g(x) = N(x \mid 0, \sigma^2)$. The variance of the importance sampling estimator is finite only in that range.

Hint: if $a > 0$, then $\int_{-\infty}^{\infty} |x|^a \exp(-bx^2) \, \mathrm{d}x < \infty$ if and only if $b > 0$.