Monte Carlo Sampling

Motivation

Monte Carlo sampling is often used in computing an approximation for the Expected value of a function $f(x)$, where $p(x)$ is the pdf of random variable X using the following equation:

$$E[f(x)] = \int f(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

1. Rejection Sampling

In rejection sampling, another density $q(x)$ is considered from which we can sample directly. Multiplying $q(x)$ by M would ensure that for all xs, $p(x) < Mq(x)$. Samples are then generated from a 2-D distribution where $X \sim q(x)$ and $U \sim U(0, Mq(x))$. The samples which satisfy $u_i > p(x_i)$ are then rejected. Equivalently, we can generate samples from a 2-D distribution where $X \sim q(x)$ and $U \sim U(0,1)$, and the samples which satisfy $u_i > \frac{p(x_i)}{Mq(x_i)}$ are then rejected.

```
i ← 0
while i ≠ N do
    x^(i) ~ q(x)
    u ~ U(0, 1)
    if u < p(x^(i)) / (M q(x^(i))) then
        accept x^(i)
        i ← i + 1
    else
        reject x^(i)
    end if
end while
```

2. Importance Sampling

There are several applications where we want to estimate

$$\theta = E[f(x)] = \int f(x)p(x)dx$$

we can re-write the above equation as

$$\theta = E[f(x)] = \int f(x)\frac{p(x)}{q(x)}q(x)dx = E[f(x)w(x)]$$

where $w(x) = \frac{p(x)}{q(x)}$.

Therefore θ can be numerically estimated as

$$\theta \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)\, w(x_i)$$

Example

Assume we want to estimate the probability $P(X > 5)$ where X has a Cauchy distribution:

$$f(x) = \frac{1}{\pi\,(1 + x^2)} \qquad -\infty < x < +\infty$$

We therefore want to find

$$\theta = \int_5^\infty f(x)dx$$

The easiest method would be to simulate values from the Cauchy distribution directly and approximate $P(X > 5)$ by the proportion of the simulated values which are bigger than 5.

The problem is that the variance of this estimator is very large as in Cauchy distribution samples are rarely exceed 5 (about 6%).

For a second we assume that the question is not about to calculate a probability, but is to think of θ not as a probability, but is to calculate the area under a curve.

for large xs, $f(x)$ is approximately equal to $\frac{1}{\pi x^2}$. We therefore like to generate a probability fuction close to this simplified $f(x)$, which would be

$$q(x) = \frac{5}{x^2} \quad x > 5$$

We then can rewrite θ as

$$\theta = \int_5^\infty \frac{f(x)}{q(x)} q(x)dx$$

We can easily generate sample from $q(x)$ using inversion method from a Uniform distribution (how?).

Therefore θ can be estimated using

$$\theta \approx \frac{1}{n}\sum_{i=1}^{n} \frac{x_i^2}{5\pi\left(1 + x_i^2\right)}$$

Background to the problem

A large manufacturing plant wishes to investigate the rate at which machines breakdown each day. You may assume that these number of machine breakdowns each day follows a Poisson(λ) distribution. Furthermore, you are 80% sure that the value of λ is less than 5 and choose to use an exponential distribution as a prior for λ. The following table displays the number of breakdowns over 50 days. However, the precise number of breakdowns is only recorded if there had been 2 or more breakdowns on a given day. It is of vital importance to understand the frequency of the number of days, f_0, where there were no breakdowns.

Number of machine breakdowns (per day)	≤ 1	2	3	4	5	6	7	8	9	10
Frequency	18	13	8	3	4	3	0	0	0	1

Aims

The primary aims of this project are to understand :

1) The posterior distribution of λ.
2) The number of days out of 50, f_0, where there were no machine breakdowns.