Stochastic Simulation
Variance reduction methods

Bo Friis Nielsen

Applied Mathematics and Computer Science

Technical University of Denmark

2800 Kgs. Lyngby – Denmark

Email: bfni@dtu.dk

Explanation: What is the problem with the Pareto distribution

- Moment distributions
- For nonnegative valued random variables

$$G_j(x) = \frac{\int_0^x t^j f(t)\mathrm{d}t}{\int_0^\infty t^j f(t)\mathrm{d}t} = \frac{\int_0^x t^j f(t)\mathrm{d}t}{\mathbb{E}\left(X^j\right)}$$

The contribution to the j'th moment from values $\leq x$.

$$\int_0^x t^1 f(t)\mathrm{d}t = \int_\beta^x t\frac{k}{\beta}\left(\frac{t}{\beta}\right)^{-k-1}\mathrm{d}t = \int_\beta^x k\left(\frac{t}{\beta}\right)^{-k}\mathrm{d}t$$

$$= \beta\frac{k}{k-1}\int_\beta^x \frac{k-1}{\beta}\left(\frac{t}{\beta}\right)^{-k}\mathrm{d}t = \frac{\beta k}{k-1}\left[1-\left(\frac{x}{\beta}\right)^{-k+1}\right]$$

Explanation: What is the problem with the Pareto distribution continued

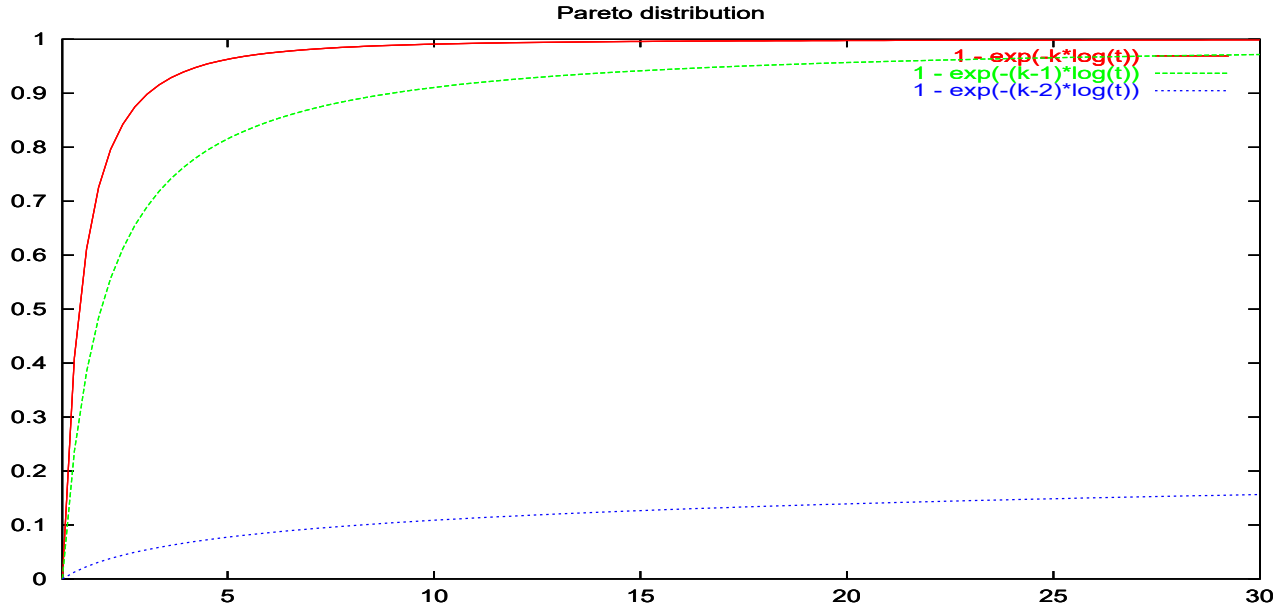
The contribution to the j'th moment from values $\leq x$.

$$\int_0^x tf(t)\mathrm{d}t = \int_\beta^x t\frac{k}{\beta}\left(\frac{t}{\beta}\right)^{-k-1}\mathrm{d}t = \frac{\beta k}{k-1}\left[1-\left(\frac{x}{\beta}\right)^{-k+1}\right]$$

$$G_1(x) = \frac{\frac{\beta k}{k-1}\left[1-\left(\frac{x}{\beta}\right)^{-k+1}\right]}{\frac{\beta k}{k-1}} = 1-\left(\frac{x}{\beta}\right)^{-k+1}$$

A Pareto distribution with the k parameter decreased by 1

Explanation: What is the problem with the Pareto distribution



- The first moment distribution for the Pareto distribution (green)
- The second moment distribution for the Pareto distribution (blue)

Some numbers $\beta = 1$

$$F(t) = 1 - t^{-k} \qquad f(t) = kt^{-k-1}$$

$$G_1(t) = 1 - t^{-k+1} \qquad G_2(t) = 1 - t^{-k+2}$$

For $k = 2.05$

t	$F(t)$	$G_1(t)$	$G_2(t)$
2	0.7585	0.5170	0.0341
10	0.9911	0.9109	0.1190
100	0.9999	0.9921	0.2057
844.5	$1 - 10^{-6}$	0.9992	0.2860

- Even when if we simulate 10^6 values we can not expect to get a decent estimate of the variance!

What to learn:

- Care is needed when using simulation
- Especially if one wants to study strange or rare phenomena.
- Always use your practical, theoretical and intuitive understanding of the system to support the analysis by simulation.

Variance reduction methods

- To obtain better estimates (tigher confidence intervals) with the same ressources
- Exploit analytical knowledge and/or correlation
- Methods:
  - ◇ Antithetic variables
  - ◇ Control variates
  - ◇ Stratified sampling
  - ◇ Importance sampling
  - ◇ Common random numbers

Case: Monte Carlo evaluation of integral

Consider the integral

$$\int_0^1 e^x \mathrm{d}x$$

We can interpret this interval as

$$\theta = \int_0^1 e^x \cdot 1 \mathrm{d}x \int_{-\infty}^{\infty} e^x \mathbf{1}_{\{x \in [0;1]\}} \mathrm{d}x = \mathsf{E}\left(e^U\right) \quad U \in \mathsf{U}(0,1)$$

To estimate the integral: sample of the random variable e^U and take the average.

$$X_i = e^{U_i} \qquad \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

This is the crude Monte Carlo estimator, "crude" because we use no refinements whatsoever.

Analytical considerations

It is straightforward to calculate the integral in this case

$$\int_0^1 e^x \mathrm{d}x = e - 1 \approx 1.72$$

The estimator X

$$\mathsf{E}(X) = e - 1 \qquad \mathsf{Var}(X) = \mathsf{E}(X^2) - \mathsf{E}(X)^2$$

$$\mathsf{E}(X^2) = \int_0^1 (e^x)^2 \, \mathrm{d}x = \frac{1}{2}\left(e^2 - 1\right)$$

Based on one observation

$$\mathsf{Var}(X) = \frac{1}{2}\left(e^2 - 1\right) - (e-1)^2 = 0.2420$$

Antithetic variables

General idea: to exploit dependnce, in particular correlation

- If the estimator is positively correlated with U_i (monotone function): Use $1 - U$ also

$$Y_i = \frac{e^{U_i} + e^{1-U_i}}{2} = \frac{e^{U_i} + \frac{e}{e^{U_i}}}{2} \qquad \bar{Y} = \frac{\sum_{i=0}^{n} Y_i}{n}$$

- The computational effort of calculating $\bar{Y}$ should be similar to the effort needed to compute $\bar{X}$.
  - ◇ By the latter expression of Y_i we can generate the same number of Y's as X's

Antithetic variables - analytical

We can analyse the example analytically due to its simplicity

$$\mathsf{E}(\bar{Y}) = \mathsf{E}(\bar{X}) = \theta$$

To calculate $\mathsf{Var}(\bar{Y})$ we start with $\mathsf{Var}(Y_i)$.

$$\mathsf{Var}(Y_i) = \frac{1}{4}\mathsf{Var}\left(e^{U_i}\right) + \frac{1}{4}\mathsf{Var}\left(e^{1-U_i}\right) + 2 \cdot \frac{1}{4}\mathsf{Cov}\left(e^{U_i}, e^{1-U_i}\right)$$

$$= \frac{1}{2}\mathsf{Var}\left(e^{U_i}\right) + \frac{1}{2}\mathsf{Cov}\left(e^{U_i}, e^{1-U_i}\right)$$

$$\mathsf{Cov}\left(e^{U_i}, e^{1-U_i}\right) = \mathsf{E}\left(e^{U_i}e^{1-U_i}\right) - \mathsf{E}\left(e^{U_i}\right)\mathsf{E}\left(e^{1-U_i}\right)$$

$$= e - (e-1)^2 = 3e - e^2 - 1 = -0.2342$$

$$\mathsf{Var}(Y_i) = \frac{1}{2}(0.2420 - 0.2342) = 0.0039$$

Comparison: Crude method vs. antithetic

Crude method:

$$\text{Var}(X_i) = \frac{1}{2}\left(e^2 - 1\right) - (e-1)^2 = 0.2420$$

Antithetic method:

$$\text{Var}(Y_i) = \frac{1}{2}(0.2420 - 0.2342) = 0.0039$$

I.e, a reduction by 98 % , almost for free.

The variance on $\bar{X}$ - and $\bar{Y}$ - will scale with $1/n$, the number of samples.

Going from crude to antithetic method, reduces the variance as much as increasing number of samples with a factor 50.

Antethetic variables in more complex models

If

$$X = h(U_1, \ldots, U_n)$$

where h is monotone in each of its coordinates, then we can use antithetic variables

$$Y = h(1 - U_1, \ldots, 1 - U_n)$$

to reduce the variance, because

$$\mathsf{Cov}(X, Y) \leq 0$$

and therefore $\mathsf{Var}(\frac{1}{2}(X + Y)) \leq \frac{1}{2}\mathsf{Var}(X)$.

Antithetic variables in the queue simulation

Can you device the queueing model of yesterday, so that the number of rejections is a monotone function of the underlying U_i's?

Yes: Make sure that we always use either U_i or $1 - U_i$, so that a large U_i implies customers arriving quickly and remaining long.

Control variates

Use of covariates

$$Z = X + c(Y - \mu_y) \qquad \mathsf{E}(Y) = \mu_y(\text{ known})$$

$$\mathsf{Var}(Z) = \mathsf{Var}(X) + c^2\mathsf{Var}(Y) + 2c\mathsf{Cov}(Y, X)$$

We can minimize $\mathsf{Var}(Z)$ by choosing

$$c = \frac{-\mathsf{Cov}(X, Y)}{\mathsf{Var}(Y)}$$

to get

$$\mathsf{Var}(Z) = \mathsf{Var}(X) - \frac{\mathsf{Cov}(X, Y)^2}{\mathsf{Var}(Y)}$$

Example

Use U as control variate

$$Z_i = X_i + c\left(U_i - \frac{1}{2}\right) \qquad X_i = e^{U_i}$$

The optimal value can be found by

$$\mathsf{Cov}(X, Y) = \mathsf{Cov}\left(U, e^U\right) = \mathsf{E}\left(Ue^U\right) - \mathsf{E}(U)\mathsf{E}\left(e^U\right) \approx 0.14086$$

In practice we would not know this covariance, but estimate it empirically.

$$\mathsf{Var}(Z_{c=\frac{-0.14086}{1/12}}) = \mathsf{Var}\left(e^U\right) - \frac{\mathsf{Cov}\left(e^U, U\right)^2}{\mathsf{Var}\left(U\right)} = 0.0039$$

Stratified sampling

This is a general survey technique: We sample in predetermined areas, using knowledge of structure of the sampling space

$$W_i = \frac{e^{\frac{U_{i,1}}{10}} + e^{\frac{1}{10}+\frac{U_{i,2}}{10}} + \cdots + e^{\frac{9}{10}+\frac{U_{i,10}}{10}}}{10}$$

What is an appropriate number of strata?

(In this case there is a simple answer; for complex problems not so)

Importance sampling

Suppose we want to evaluate

$$\theta = \mathsf{E}(h(X)) = \int h(x)f(x)\mathrm{d}x$$

For $g(x) > 0$ whenever $f(x) > 0$ this is equivalent to

$$\theta = \int \frac{h(x)f(x)}{g(x)} g(x)\mathrm{d}x = \mathsf{E}\left(\frac{h(Y)f(Y)}{g(Y)}\right)$$

where Y is distributed with density $g(y)$.

This is an efficient estimator of θ, if we have chosen g such that the variance of $\left(\frac{h(Y)f(Y)}{g(Y)}\right)$ is small.

Such a g will lead to more Y's where $h(y)$ is large.

More important regions will be sampled more often. IS can also be used, when sampling from g is easier than sampling from f, in this sense IS is closely related to acceptance/rejection.

Re-using the random numbers

We want to compare two different queueing systems.

We can estimate the rejection rate of system $i = 1, 2$ by

$$\hat{\theta}_i = \mathsf{E}(g_i(U_{i1}, \ldots, U_{in}))$$

and then rate the two systems according to

$$\hat{\theta}_2 - \hat{\theta}_1$$

But typically $g_1(U_1, \cdots, U_n)$ and $g_2(U_1, \cdots, U_n)$ are positively correlated: Long service times imply many rejections.

Then a more efficient estimator is based on

$$\hat{\theta}_2 - \hat{\theta}_1 = \mathsf{E}\left(g_2(U_1, \ldots, U_n) - g_1(U_1, \ldots, U_n)\right)$$

This amounts to letting the two systems run with the *same* input sequence of random numbers, i.e. same arrival and service time for each customer. However, care is needed, different customers could be blocked in the two systems.

With some program flows, this is easily obtained by re-setting the seed of the RNG.

When this is not sufficient, you must store the sequence of arrival and service times, so they can be re-used.

Exercise 5: Variance reduction methods

1. Estimate the integral $\int_0^1 e^x dx$ by simulation (the crude Monte Carlo estimator). Use eg. an estimator based on 100 samples and present the result as the point estimator and a confidence interval.
2. Estimate the integral $\int_0^1 e^x dx$ using antithetic variables, with comparable computer ressources.
3. Estimate the integral $\int_0^1 e^x dx$ using a control variable, with comparable computer ressources.
4. Estimate the integral $\int_0^1 e^x dx$ using stratified sampling, with comparable computer ressources.
5. Use control variates to reduce the variance of the estimator in exercise 4 (Poisson arrivals).
6. Demonstrate the effect of using common random numbers in exercise 4 for the difference between Poisson arrivals (Part 1) and a renewal process with hyperexponential interarrival times. **Remark:** You might need to do some thinking and some re-programming.

7. For a standard normal random variable $Z \sim \mathrm{N}(0,1)$ using the crude Monte Carlo estimator estimate the probability $Z > a$. Then try importance sampling with a normal density with mean a and variance σ^2. For the expirements start using $\sigma^2 = 1$, use different values of a (e.g. 2 and 4), and different sample sizes. If time permits experiment with other values for σ^2. Finally discuss the efficiency of the methods.
8. Use importance sampling with $g(x) = \lambda \exp(-\lambda * x)$ to calculate the integral $\int_0^1 e^x \mathrm{d}x$ of Question 1. Try to find the optimal value of λ by calculating the variance of $h(X)f(X)/g(X)$ and verify by simulation. Note that importance sampling with the exponential distribution will not reduce the variance.
9. For the Pareto case derive the IS estimator for the mean using the first moment distribution as sampling distribution. Is the approach meaningful? and could this be done in general? With this insight could you change the choice of $g(x)$ in the previous question

(Question 8) such that importance sampling would reduce the variance? You do not need to implement this, as long as you can argue, what should happen.