Monte Carlo Methods and Simulations in Nuclear Technology

RNG, sampling procedures, simple sampling

Jan Dufek

2022

KTH Royal Institute of Technology

Topics

Topics

- RNG
- Sampling by the inverse transform method
- Sampling by the acceptance-rejection method
- The central limit theorem
- Mean value of the result, variance of the mean value of the result
- Variance of the mean vs the number of samples
- Confidence intervals
- Accuracy vs precision

Generators of random numbers (RNG)

Generators of random numbers (RNG)

What is the purpose of RNGs?

The purpose of RNGs is to generate numbers randomly from a uniform distribution between 0 and 1, $\mathscr{U}(0,1)$. RNGs are used in Monte Carlo simulations to sample values of random variables.

What qualities the RNGs should have?

- Randomness
- Reproducibility
- Large length of the sequence of random numbers
- Reasonable computer memory demands
- Small generation time

Generators of random numbers (RNG)

Name some types of RNGs

The most convenient and reliable way of generating the random numbers for simulations is via deterministic algorithms - pseudo-random number generators, such as:

- *linear congruential generator* (LCG)
- *multiple recursive generator* (MRG)
- *nonlinear generators*

How does the linear congruential generator (LCG) work?

- Integer numbers, x_n, are generated by the recurrence

$$x_n = (ax_{n-1} + c) \bmod m,$$

  where

  - $m > 0$ is the modulus (must be an integer number),
  - $a > 0$ is the multiplier (must be an integer number),
  - c is the additive constant (must be an integer number).

- The "mod m" is the operation of taking the least nonnegative residue modulo m.
- In order to produce values u_n in the interval [0, 1], the value x_n must be divided by m,

$$u_n = x_n/m.$$

- In C language, you'd have to write this operation as

```
u = float(x) / float(m)
```

- The maximal period length for LCGs is m.
- The x_0 is called the **seed**, and it can be set arbitrary.

Sampling procedures

Sampling procedures

Objective of sampling procedures

- The objective is to generate samples $x_1, x_2, \ldots$ of a random variable X from a given pdf $f_X(x)$ or cdf $F_X(x)$.
- All sampling procedures generate samples x from f_X or F_X by transforming random numbers generated by RNG from pdf $\mathscr{U}(0, 1)$.

How does the inverse transform method work?

The *inverse transform method* provides the most direct way of generating samples from $F_X(x)$ on the interval $[a, b]$. It uses the inverted form of $F_X(x)$. A random sample x of X can be then obtained easily as

$$x = F_X^{-1}(u),$$

where u is randomly sampled from pdf $\mathscr{U}(0, 1)$.

Does the inverse transform method have some disadvantage?

- It may not be possible to invert the cumulative distribution function $F_X(x)$.
- Even when F_X^{-1} exists for a given random variable X, it *may* not be in a form suitable for efficient computation.

How does the acceptance-rejection method work?

- This technique generates samples from any pdf $f_X(x)$ using another pdf $h(x)$ for that holds that

$$f_X(x) \leq h(x)c,$$

  where $c = \sup_x[f_X(x)/h(x)]$. (Note that c is ≥ 1.)

- Procedure to generate one sample:
  - generate two random numbers:
    - x from $h(x)$, and
    - u from $\mathscr{U}(0, 1)$
  - accept x if

$$u \times c \times h(x) < f_X(x),$$

    else reject x, and start from the beginning.

What can we say about the efficiency of the acceptance-rejection method?

- The proportion of proposed samples which are accepted is

$$\frac{\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x}{\int_{-\infty}^{\infty} c \times h(x)\mathrm{d}x} = \frac{1}{c}$$

- Thus, to ensure a good efficiency, c should be close to unity.
- This can be satisfied only when $h(x)$ is chosen close to $f_X(x)$.
- Yet, it must be possible to generate the samples from $h(x)$ easily using the inverse transform method.

Simple sampling

Simplest form of a Monte Carlo calculation

- In its simplest form, a Monte Carlo calculation generates samples of an (input) known random variable X (with a known pdf f_X), and samples an unknown random variable Y using a numerical model (function) g,

$$Y = g(X)$$

- The objective of the Monte Carlo calculation is to estimate the expectation value and variance of Y, and the variance of m_Y (where m_Y is the average value of collected samples of Y).
- Example:
    - X ... kinetic energy of fission neutrons
    - Y ... distance to the first collision

Simple sampling

Having sampled n values, y_1, y_2, . . ., y_n of Y,

we can **estimate** the expectation value of Y by the mean value

$$m_Y = \frac{1}{n}\sum_{i=1}^{n} y_i$$

What does the central limit theorem say about relation of m_Y and $\mathrm{E}[Y]$?

The CLT says that the mean value m_Y is a random variable and that m_Y is normally distributed with a mean $\mathrm{E}[Y]$,

$$\mathrm{E}[m_Y] = \mathrm{E}[Y].$$

Simple sampling

Variance of the mean

The variance of m_Y, which estimates the precision of the computed m_Y, equals

$$\begin{aligned}\mathrm{Var}[m_Y] &= \mathrm{E}[(m_Y - \mathrm{E}[Y])^2] = \\ &= \mathrm{E}\left[\left(\frac{\sum y_i}{n} - \mathrm{E}[Y]\right)^2\right] = \mathrm{E}\left[\left(\frac{\sum(y_i - \mathrm{E}[Y])}{n}\right)^2\right] = \\ &= \mathrm{E}\left[\frac{(\sum \xi_i)^2}{n^2}\right] = \frac{\mathrm{E}(\sum \xi_i)^2}{n^2} = \\ &= \frac{1}{n^2}\left[\sum \mathrm{E}[\xi_i^2] + 2\sum \mathrm{E}[\xi_i \xi_{i+1}] + 2\sum \mathrm{E}[\xi_i \xi_{i+2}] + \dots\right],\end{aligned}$$

where

$$\xi_i \equiv y_i - \mathrm{E}[Y].$$

Variance of the mean (cont.)

If ξ_i are statistically independent then the cross products $\mathrm{E}[\xi_i \xi_{i+1}]$, $\mathrm{E}[\xi_i \xi_{i+2}]$, etc. in previous equation equal zeros, and

$$\begin{aligned}\mathrm{Var}[m_Y] &= \frac{\sum \mathrm{E}[\xi_i^2]}{n^2} \\ &= \frac{n\mathrm{E}[\xi_i^2]}{n^2} \\ &= \frac{\mathrm{Var}[Y]}{n}.\end{aligned}$$

How does the variance of the mean value converge with the number of samples?

Nevertheless, $\mathrm{Var}[Y]$ above is not known, and must be estimated by

$$s_Y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - m_Y^2$$

i.e., the variance of m_Y is approximated by

$$s_{m_Y}^2 = \frac{s_Y^2}{n}$$

Simple sampling

Will the error in m_Y, i.e. the difference $m_Y - \mathrm{E}[Y]$, decrease with collecting more samples y_i?

When collecting more samples y_i the **estimated** $\mathrm{Var}[m_Y]$ will *usually* decrease; however, the **real** error in m_Y is never known and it may even increase when more samples are collected.

Accuracy vs precision

- Accuracy of the calculation reflects the error in m_Y, i.e. the difference $m_Y - \mathrm{E}[Y]$. An accurate calculation returns a result with a small error (relative to the result value).
- Precision of the calculation reflects the variance of the mean value of the result. A precise calculation has a small standard deviation of the mean value relative to the mean value of the result.
- A precise calculation may not necessarily be accurate (e.g. when incorrect input data cause a large error).
- An accurate calculation may not necessarily be precise (e.g., when a small error in the result is achieved by a chance).

Confidence intervals

Confidence intervals

Can we estimate the probability that $\mathrm{E}[Y]$ lies within a certain confidence interval $(m_Y - \delta, m_Y + \delta)$ for an arbitrarily chosen δ? How?

The probability that $\mathrm{E}[Y]$ is inside the interval $[m_Y - \delta, m_Y + \delta]$ equals probability that m_Y is inside $(\mathrm{E}[Y] - \delta, \mathrm{E}[Y] + \delta)$:

$$
\begin{aligned}
P &= \int_{\mathrm{E}[Y]-\delta}^{\mathrm{E}[Y]+\delta} \frac{1}{\sigma_{m_Y}\sqrt{2\pi}} \exp\left(-\frac{(y - \mathrm{E}[Y])^2}{2\sigma_{m_Y}^2}\right) dy \\
&= \frac{2}{\sigma_{m_Y}\sqrt{2\pi}} \int_{\mathrm{E}[Y]}^{\mathrm{E}[Y]+\delta} \exp\left(-\frac{(y - \mathrm{E}[Y])^2}{2\sigma_{m_Y}^2}\right) dy \\
&= \frac{2}{\sigma_{m_Y}\sqrt{2\pi}} \int_0^{\delta} \exp\left(-\frac{y^2}{2\sigma_{m_Y}^2}\right) dy \\
&= \mathrm{erf}\left(\frac{\delta}{\sigma_{m_Y}\sqrt{2}}\right),
\end{aligned}
$$

where $\mathrm{erf}(x)$ is the Gauss error function. Since σ_{m_Y} in is not known, it must be approximated by s_{m_Y}.

Probability level P for various confidence intervals

The probability

$$P = \mathrm{erf}\left(\frac{\delta}{\sigma_{m_Y}\sqrt{2}}\right),$$

where $\mathrm{erf}(x)$ is the Gauss error function, is tabulated for a number of values. Since σ_{m_Y} in is not known, it must be approximated by s_{m_Y}.

P	δ/σ_{m_Y}
0.6826895	1
0.9544997	2
0.9973002	3
0.9999366	4
0.9999994	5

Confidence intervals

In which form the result of a Monte Carlo simulation is usually written?

- The result can be written as

$$m_Y \pm s_{m_Y}$$

where s_{m_Y} is the estimation of the standard deviation of the mean value m_Y,

$$s_{m_Y} = \frac{s_Y}{\sqrt{n}}$$

where s_Y is the estimation of the standard deviation of random variable Y,

$$s_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - m_Y^2}$$

- When the result is given simply as $m_Y \pm s_{m_Y}$ then it is assumed that the result is inside *one-sigma* interval $(m_Y - \delta, m_Y + \delta)$ with $\delta = s_{m_Y}$ with probability about 0.68.
- When another confidence interval is used then it must me specified as to how many *sigmas* the confidence interval $m_Y \pm s_{m_Y}$ has.