2. Monte Carlo Simulation

All the methods we will use in this course will assume the availability of a sequence of random numbers $u_1, u_2, \ldots$ which are independent and identically distributed (i.i.d.). We will first discuss how to generate uniformly distributed sequences, and then how to use these samples to generate non-uniform random numbers.

2.1. Generating uniform random numbers.

There are two main methods to generate i.i.d. random variables (r.v.) which are uniformly distributed in $[0,1]$ ($\mathcal{U}(0,1)$ distributed).

(1) Involves some physical phenomenon which is expected to be random. One of these approaches involves measuring radioactive decay: the times between successive decay events (from a radioactive source) are known to be i.i.d. exponentially distributed r.v., which can be transformed to $\mathcal{U}(0,1)$. Another example would be to measure atmospheric noise, etc. Disadvantages: not very fast, require measurement equipment, AND are not reproducible.

Because of these disadvantages, in practice we use

(2) Pseudorandom number generators (PRNG)

This is the standard way in which computers produce random numbers nowadays. It consists of using a deterministic algorithm to produce a sequence of numbers which is "very close to being random", i.e., it passes a stringent number of statistical tests.

Definition: A pseudorandom number generator is a recursive algorithm which, given an inital seed x_0, produces a sequence $u_1, u_2, \ldots \in [0,1]$ constructed by

$$u_i = g(x_i), \qquad x_i = f(x_{i-1}), \quad i \geq 1$$

- Note that if we use the same value x_0, then the algorithm produces the same sequence $\Rightarrow$ sequence is reproducible!
- The set of possible values of $(x_i)_{i \in \mathbb{N}}$ is finite $\Rightarrow$ the PRNG will eventually repeat. In this case, if we have $x_{\ell+d} = x_\ell$. The smallest d for which this occurs is called the period of the PRNG.

Example: A simple example of a PRNG is the linear congruential generator (LCG):

$$u_{n+1} = \frac{x_{n+1}}{M}, \text{ where } x_{n+1} = (a x_n + c) \bmod M.$$

- the period of the LCG is less than or equal to M, and it is less in general, depending on a and c.
- the Hull-Dobell theorem provides necessary and sufficient conditions for a, c and M to satisfy in order for the LCG to have period M for all seeds x_0. For example, glibc's rand() uses
$$M = 2^{32}, \quad a = 22695477, \quad c = 1$$
- This method is not used in practice, as it does not produce sufficiently independent samples. Another example is the middle-square algorithm, or the Mersenne Twister, which is used by Matlab, Julia, etc.

For the remainder of the module, we will assume that we are provided with a sequence of iid uniformly distributed random numbers.

Testing the uniform distribution hypothesis

Several tests can be employed to assess random or pseudorandom number generators: the Kolmogorov–Smirnov (KS) test, the χ^2 test, the Diehard tests, and many more. A detailed discussion of theses tests is beyond the scope of this course, so here we include only a brief discussion of the KS test for the sake of illustration. We will then employ the KS test to assess the performance of the Linear Congruential Generator (LCG) with the parameters employed in the `glibc` library.

The KS test is based on a comparison between the expected cumulative distribution function (CDF) and an empirical CDF constructed from the numbers given by our generator. Let us assume that $\{X_i\}_{i\in\mathbb{N}}$ is a sequence of independent, identically distributed (i.i.d.) real-valued random variables with CDF F, and let us define the empirical CDF associated to $\{X_i\}_{i=1}^n$ by

$$F_N(x) = \frac{1}{N}\sum_{i=1}^{N} I_{(-\infty,x]}(X_i),$$

where I_A denotes the indicator function of the set A. By the strong law of large numbers (SLLN), for all $x \in \mathbb{R}$ it holds that

$$F_N(x) = \frac{1}{N}\sum_{i=1}^{N} I_{(-\infty,x]}(X_i) \to \mathbb{E}\left[I_{(-\infty,x]}(X_i)\right] = \mathbb{P}\left[X_i \le x\right] = F(x) \quad \text{a.s. when } N \to \infty. \tag{1}$$

This means that for all $x \in \mathbb{R}$, the event $A_x := \{\lim_{N\to\infty} F_N(x) = F(x)\}$ has probability one. A good PRNG should therefore produce numbers such that, for any x, the empirical CDF at x converges to $F(x)$ with probability 1 as the number of samples increases.

Remark 1. Equation (1) is not sufficient to deduce that F_N converges pointwise (in x) to F almost surely. Indeed, let us denote by N_x the complement (in the underlying sample space Ω) of A_x, i.e. N_x is the event that $F_N(x)$ does not converge to $F(x)$. The event "$\lim_{N\to\infty} F_N(x) = F(x)$ for all $x \in \mathbb{R}$", which we denote by B, has probability $\mathbb{P}[B] = 1 - \mathbb{P}[B^c] = 1 - \mathbb{P}\left[\bigcup_{x\in\mathbb{R}} N_x\right]$, and since an uncountable union of events with probability 0 does not necessarily have probability zero, we cannot conclude that $\mathbb{P}[B] = 1$. ⊘

Fortunately, the following theorem shows that, in fact, F_N does converge to F almost surely, pointwise (in x) and even uniformly.

Theorem 1 (Glivenko–Cantelli)**.** *With the same notations and assumptions as above,*

$$D_N := \sup_{x\in\mathbb{R}} |F_N(x) - F(x)| \to 0 \quad \textit{a.s. when } N \to \infty.$$

Proof. For simplicity, we consider only the case where F is continuous. Let $-\infty = x_0 < x_1 < \cdots < x_m = +\infty$ be such that $F(x_i) = \frac{1}{m}$ (this is possible because F is continuous). For any $x \in \mathbb{R}$, there exists $j \in \{0, \dots, m-1\}$ such that $x \in [x_j, x_{j+1}]$. Since both F and F_N are nondecreasing,

$$\begin{cases} F_N(x) - F(x) \le F_N(x_{j+1}) - F(x_j) = F_N(x_{j+1}) - F(x_{j+1}) + \dfrac{1}{m}, \\ F_N(x) - F(x) \ge F_N(x_j) - F(x_{j+1}) = F_N(x_j) - F(x_j) - \dfrac{1}{m}. \end{cases} \tag{2}$$

Let now $M_N := \max_{j\in\{0,\dots,m\}} |F_N(x_j) - F(x_j)|$. It follows from Eq. (2) and the fact that x was arbitrary that

$$D_N = \sup_{x\in\mathbb{R}} |F_N(x) - F(x)| \leq M_N + \frac{1}{m}.$$

Since M_N, being the maximum of a finite number of random variables that converge a.s. to 0 as $N \to \infty$, also converges a.s. to 0 as $N \to \infty$, we deduce from the previous inequality that

$$L := \limsup_{N\to\infty} D_N \leq \frac{1}{m} \quad \text{a.s.}$$

To conclude, note that $\{L > 0\} = \bigcup_{m=1}^{\infty} \{L > \frac{1}{m}\}$, so by countable subadditivity

$$\mathbb{P}[L > 0] \leq \sum_{i=1}^{\infty} \mathbb{P}\left[L > \frac{1}{m}\right] = 0,$$

i.e. $L = 0$ a.s. □

With this result, we can refine our expectations of a good random number generator: a good generator should produce samples such that the empirical CDF converges uniformly to F with probability 1. The Glivenko–Cantelli theorem, however, is of little practical use for us, because it does not contain any information on the speed of convergence of D_n to 0. To build the KS test statistic, we need the following stronger result:

Theorem 2 (Kolmogorov)**.** *With the same notations and assumptions as above,*

$$\sqrt{N}\, D_N \xrightarrow{d} K,$$

where K is the Kolmogorov distribution, with CDF

$$\mathbb{P}[K \leq x] = 1 - 2\sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2}.$$

In its simplest form, the Kolmogorov–Smirnov test is based on the approximation that $\sqrt{N}D_N$ follows the Kolmogorov distribution even when N is finite: For large enough N,

$$\mathbb{P}[\sqrt{N}\, D_N \geq \varepsilon] \approx \lim_{N\to\infty} \mathbb{P}[\sqrt{N}\, D_N \geq \varepsilon] = \mathbb{P}[K \geq \varepsilon]. \tag{3}$$

Remark 2. We emphasize that, while it is necessary for a good PRNG to pass the KS test, it is far from sufficient, and in practice many other tests should be employed before coming to a conclusion. Consider the following example:

```
import numpy as np
import scipy.stats as stats
x = np.linspace(0, 1, 10**6)
print(stats.kstest(x, 'uniform'))
```

Our sequence is clearly not random, yet SciPy returns a p-value of 1. In this context, the p-value is the approximate probability in Eq. (3). ⊘

2.2. Generating non-uniform random numbers

2.2.1. Inverse transform method

Suppose we want to produce samples from a r.v. X which has cumulative distribution function (CDF) $F(x)$ (i.e., $F(x) = \mathbb{P}(X \leq x)$). If we have access to a PRNG that generates samples from $\mathcal{U}(0,1)$ then for one-dimensional distributions we can use these to generate samples from X.

If F is continuous and strictly increasing, we define $G(u) = F^{-1}(u)$, i.e., $x = G(u)$ is the unique solution to $F(x) = u$. If F is discontinuous or not strictly increasing, then

$$G(u) := \inf\{x : F(x) \geq u\},$$

and if $0 < u < 1$ we still have $F(G(u)) = u$.

Lemma: If $U \sim \mathcal{U}(0,1)$ and F is a one-dimensional CDF, then $X = G(U)$ has CDF F.

Proof: We can check that $G(u) \leq x \iff u \leq F(x)$. Therefore, $\mathbb{P}(G(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$. ∎

The Inverse Transform method

1. Generate a random number u from $\mathcal{U}(0,1)$
2. Compute $x = G(u)$
3. Take x to be a sample of the r.v. X with CDF F.

example: Suppose we wish to sample from $X \sim \text{Exp}(\lambda)$ (exponential distribution with rate λ). The CDF of X is $F(x) = 1 - e^{-\lambda x}$. We can compute $G(u) = F^{-1}(u)$:

$$G(u) = -\frac{1}{\lambda}\ln(1-u).$$

$\Rightarrow X := -\frac{1}{\lambda}\ln(1-U)$, $U \sim \mathcal{U}(0,1)$ is distributed according to f!

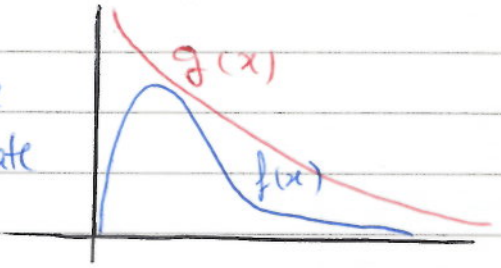
Note: Since $1 - U \sim \mathcal{U}(0,1)$, we can also use $X = -\frac{1}{\lambda}\ln U$.

2.2.2. Rejection Sampling

Suppose now that we wish to generate samples of a r.v. X with a known and computable density $f(x)$. If there is a density $g(x)$ on the same state space E, from which it is easy to generate samples and

$$f(x) \leq M g(x) \quad \forall x \in E$$

holds for some finite M, then we can use rejection sampling to generate samples of f from samples of g.



We generate a sample x from $g(x)$, and accept $X = x$ with probability $f(x)/Mg(x)$, otherwise reject it and repeat this process until accepted.

Rejection Sampling Method:

1. Generate a sample x from $g(x)$ and let $u \sim U(0,1)$
2. If $u < f(x)/Mg(x)$, accept $X = x$ and stop
3. Otherwise, reject and return to (1)

Note that the rejection sampling method can be used to sample from continuous AND discrete distributions → replace pdf by pmf.

Lemma: Let Z be the r.v. generated by the rejection sampling method. Then Z is distributed according to f.

Proof: W.l.o.g. assume $E = \mathbb{R}$. Let Y be a r.v. with density g and denote by $A = 1$ the event that the sample is accepted, i.e., $U < f(Y)/Mg(Y)$, $U \sim U(0,1)$, $Y \sim g$.

$$P(A=1) = \int_{-\infty}^{+\infty} P(A=1 \mid Y=y)\, g(y)\, dy$$

$$= \int_{-\infty}^{+\infty} g(y) \frac{f(y)}{Mg(y)}\, dy = \frac{1}{M} \int_{-\infty}^{+\infty} f(y)\, dy = 1/M.$$

Furthermore, for $r \in \mathbb{R}$ fixed,

$$\mathbb{P}(Y \le r, A = 1) = \int_{-\infty}^{r} \left(\int_0^{f(x)/Mg(x)} du \right) g(x)\,dx = \frac{1}{M} \int_{-\infty}^{r} f(x)\,dx$$

and therefore

$$\mathbb{P}(Z \le r) = \mathbb{P}(Y \le r \mid A = 1) = \frac{\mathbb{P}(Y \le r, A = 1)}{\mathbb{P}(A = 1)} = \int_{-\infty}^{r} f(x)\,dx$$

and $Z \sim f$. ∎

Note that since f and g integrate to 1, $M \ge 1$. For efficiency, we want to reject as few samples as possible. The probability of acceptance is $\frac{1}{M} \Rightarrow$ Want M as close to 1 as possible $\rightarrow$ want g as close to f as possible.

Note that in many cases from applications, we know f (or g) up to some normalising constant, i.e.,

$$\int f(x)\,dx = Z \neq 1, \qquad \int g(x)\,dx = Z' \neq 1.$$

In this case, we can still apply rejection sampling (i.e., no need to compute these integrals!).

Let $\tilde{f} = f/Z$, $\tilde{g} = g/Z'$. If $f(x) \le Mg(x)$, we still have

$$\tilde{f}(x) = f(x)/Z \le Mg(x)/Z = \frac{MZ'}{Z} \frac{g(x)}{Z'} = M' \tilde{g}(x), \quad x \in \mathbb{R}.$$

So we can perform rejection sampling with $\tilde{f}$ and $\tilde{g}$ instead: we accept a sample $x \sim \tilde{g}$ if

$$u \le \frac{\tilde{f}(x)}{M' \tilde{g}(x)} \quad (=) \quad u \le \frac{f}{Z} \cdot \frac{Z'}{g} \cdot \frac{Z}{MZ'} = \frac{f}{Mg}$$

where $u \sim \mathcal{U}(0,1)$.

$\Rightarrow$ We can safely ignore normalising constants from the beginning!

In this case, the probability of accepting a proposal is Z/MZ'.

2.2.3. Sampling from Gaussian distributions

There is no closed formula for the CDF of a Gaussian distribution $f(x) = 1/2\pi \int_{-\infty}^{x} e^{-y^2/2} dy \Rightarrow$ we cannot use the inverse transform method. One could use rejection sampling with a Cauchy proposal distribution (exercise). However, by applying the right transformation, we can obtain a pair (X, Y) of iid standard Gaussian r.v. directly from a pair of iid $U(0,1)$ r.v.s

This method is called the Box-Muller algorithm. It uses the transformation of the PDF of a pair of independent standard normals (X, Y):

$$f(x,y) = 1/2\pi \; e^{-(x^2+y^2)/2}$$

into polar coordinates: $(X, Y) = R(\cos\Theta, \sin\Theta)$, where $R > 0$, $0 \le \Theta < 2\pi$. We have:

- $\Theta \sim U(0, 2\pi)$ because (X, Y) is rotationally symmetric around the origin $\Rightarrow \Theta = 2\pi U_1$, $U_1 \sim U(0,1)$
- $P(R \le r) = 1/2\pi \int_0^r \int_0^{2\pi} e^{-r'^2/2} \, r' \, d\theta \, dr'$
  $= \int_0^r \exp(-r'^2/2) \, r' \, dr' = 1 - e^{-r^2/2}$.

We can invert $F(r) = 1 - e^{-r^2/2} \Rightarrow R = \sqrt{-2\ln(U_2)}$, $U_2 \sim U(0,1)$

$\Rightarrow \boxed{(X, Y) = \left(\sqrt{-2\ln U_2}\cos(2\pi U_1), \sqrt{-2\ln(U_2)}\sin(2\pi U_1)\right)}$

Once we have $X \sim \mathcal{N}(0,1)$, it is straightforward to generate $Y \sim \mathcal{N}(\mu, \sigma^2)$: $Y = \mu + \sigma X$.

The proof that (X, Y) has the right distribution follows from the fact that if

$$x = \sqrt{-2\ln u_2}\cos(2\pi u_1), \quad y = \sqrt{-2\ln u_2}\sin(2\pi u_1)$$

then

$$u_1 = e^{-(x^2+y^2)/2}, \quad u_2 = 1/2\pi \tan^{-1}(y/x).$$

2.2.4. Multivariate Gaussian Distributions

The Box-Muller algorithm allows us to sample from uncorrelated Gaussian r.v.. But in the general case, these variables might be correlated.

Definition: Let $m \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. The r.v. $X: \Omega \to \mathbb{R}^n$ with PDF

$$\gamma_{\Sigma,m}(x) := ((2\pi)^n \det \Sigma)^{-1/2} \exp\left(-\frac{1}{2}\langle \Sigma^{-1}(x-m), (x-m)\rangle\right)$$

is a multivariate Gaussian or normal r.v.. Its mean is $\mathbb{E}(X) = m$ and covariance matrix $\mathbb{E}((X-m) \otimes (X-m)) = \Sigma$. Here,
$\langle \cdot, \cdot \rangle \to$ standard Euclidean inner product
$(A \otimes B)_{ij} = A_i B_j$, $i, j \in \{1, \dots, n\}$.
We write $X \sim \mathcal{N}(m, \Sigma)$.

If $\underline{Y} \sim \mathcal{N}(0, I_{n \times n})$, i.e., $\underline{Y} = (Y_1, \dots, Y_n)$ with $Y_1, \dots, Y_n \sim \mathcal{N}(0,1)$ iid (which we can obtain from Box-Muller) and C is a matrix such that $CC^T = \Sigma$, then $\underline{X} = m + CY$ is such that $\underline{X} \sim \mathcal{N}(m, \Sigma)$.

So, all we need to generate multivariate Gaussians is the matrix C such that $CC^T = \Sigma$. A natural choice is the matrix square root of Σ. We can compute C by

- diagonalising Σ: $\Sigma = BDB^T$, B orthogonal, D diagonal
  $\Rightarrow C = B\sqrt{D}B^T$.
- Using Cholesky decomposition: $\Sigma = LL^T$ with L lower triangular (computational cost $O(n^3)$!)

2.3. Monte Carlo Simulation

We can now consider the first example described in the introduction. Given a r.v. X with density $\pi(x)$, we want to estimate integrals of the form

$$I = \mathbb{E}_{X\sim\pi}(f(X)) = \int f(x)\,\pi(x)\,dx.$$

As described in the introduction, if we can produce a sequence $x_1, x_2, \dots$ of iid samples of π, we can approximate I using the estimator

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i).$$

In this section, we will study the properties of the estimator $\hat{I}_n$. In order to do that, we need to recall the following limit theorems for iid sequences of r.v.:

Theorem (Strong Law of Large Numbers): Let $\{Z_i\}_{i\in\mathbb{N}}$ be a sequence of iid integrable r.v. with $\mathbb{E}(Z_i) = \mu$ and consider

$$S_n := \frac{1}{n}\sum_{i=1}^{n} Z_i.$$

Then S_n converges to μ almost surely, i.e.,

$$\mathbb{P}\left(\lim_{n\to\infty} S_n = \mu\right) = 1.$$

Theorem (Central Limit Theorem): Let $\{Z_i\}_{i\in\mathbb{N}}$ be a sequence of iid, square integrable (i.e., $\mathbb{E}|Z_i|^2 < \infty$) r.v. with $\mathbb{E}(Z_i) = \mu$, $\mathrm{Var}(Z_i) = \sigma^2$. Then

$$\sqrt{n}\,(S_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

i.e.

$$\lim_{n\to\infty} \mathbb{P}\left(a < \frac{\sum_{i=1}^{n} Z_i - n\mu}{\sigma\sqrt{n}} < b\right) = \frac{1}{\sqrt{2\pi}}\int_a^b e^{-x^2/2}\,dx \qquad \forall a \le b \in \mathbb{R}.$$

- The Strong LLN provides us with information about the behaviour of a sum of r.v. (or a large number of repetitions of the same experiment) on average. S_n converges almost surely (a.s.) to μ, but the estimate fluctuates around the average.
- The CLT allows us to quantify these fluctuations.
- Both theorems have been widely studied and the iid assumption can be relaxed.

We can now study the properties of the MC estimator $\hat{I}_n$.

Definition: A family of estimators $(\hat{\theta}_n)_{n\in\mathbb{N}}$ for θ is said to be:

(a) unbiased if $\mathbb{E}(\hat{\theta}_n) = \theta$, $\forall n \in \mathbb{N}$

(b) asymptotically unbiased if $\mathbb{E}(\hat{\theta}_n) \longrightarrow \theta$ as $n \to \infty$.

(c) weakly consistent if

$$\lim_{n\to\infty} \hat{\theta}_n = \theta \text{ in probability, i.e.,}$$

$$\forall a > 0 \quad \mathbb{P}(|\hat{\theta}_n - \theta| > a) \longrightarrow 0 \text{ as } n \to \infty$$

(d) strongly consistent if

$$\lim_{n\to\infty} \hat{\theta}_n = \theta \text{ almost surely}$$

(e) asymptotically normal if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \text{ as } n \to \infty.$$

Clearly a family of unbiased estimators is asymptotically unbiased. Similarly, if it is strongly consistent, it is also weakly consistent.
We will now demonstrate that all these properties hold for the MC estimator.