# CLASS NOTES

# Models, Algorithms and Data: Introduction to computing

# 2019

Petros Koumoutsakos, Jens Honore Walther, Julija Zavadlav

(Last update: July 15, 2019)

## IMPORTANT DISCLAIMERS

Much of the material (ideas, definitions, concepts, examples, etc) in these notes is taken for teaching purposes, from several references:

- Numerical Analysis by R. L. Burden and J. D. Faires
- The Nature of Mathematical Modeling by N. Gershenfeld
- A First Course in Numerical methods by U. M. Ascher and C. Greif

These notes are only informally distributed and intended ONLY as study aid for the final exam of ETHZ students that were registered for the course Models, Algorithms and Data (MAD): Introduction to computing 2019. The notes have been checked, however they may still contain errors so use with care.

# LECTURE 11 Numerical integration IV: Monte Carlo

In this chapter, we discuss another method where the locations of the data points (abscissas) are not uniform. The motivation to introduce another method, which is different from the already "optimal" Gauss quadrature will become clear when going to high dimensional integrals. An example of a high-dimensional integration is, for instance, a statistical mechanics model with $N$ particles, where $6N$-dimensional integrals ($3N$ positions and $3N$ momenta) need to be estimated. As an example considering that a liter of water contains $N \sim 10^{26}$ water molecules, these integrals are very high dimensional. As we will see in the following section for such integrals we run into problem. Luckily we can resolve these problems by means of Monte Carlo integration.

## 11.1 Curse of dimensionality

For a given multi-variate function depending on $d \in \mathbb{N}$ variables,

$$f : \mathbb{R}^d \to \mathbb{R}, \qquad \vec{x} = (x^{(1)}, \ldots, x^{(d)}) \mapsto f(x^{(1)}, \ldots, x^{(d)}), \tag{11.1.1}$$

we are interested in estimating its integral over a multi-dimensional domain $\Omega = \Omega_1 \times \cdots \times \Omega_d$, where $\Omega_r = [a_r, b_r]$ for $r = 1, \ldots, d$

$$I = \int_\Omega f(x^{(1)}, \ldots, x^{(d)}) \mathrm{d}\vec{x} \tag{11.1.2}$$

By Fubinis theorem one way to estimate $I$ would be to split it into $d$ nested integrals

$$I = \int_{\Omega_1} \cdots \int_{\Omega_d} f(x^{(1)}, \ldots, x^{(d)}) \mathrm{d}x^{(d)} \cdots \mathrm{d}x^{(1)}, \tag{11.1.3}$$

and to apply any of the 1-dimensional quadrature rules with $n$ points from the previous sections along each dimension $r = 1, \ldots, d$. In particular, given a one dimensional quadrature rule with locations $x_1, \ldots, x_n$ and weights $w_1, \ldots, w_n$, the $d$-dimensional quadrature rule obtained using this Cartesian product is given by

$$I \approx \sum_{\substack{i_1=1 \\ \cdots \\ i_d=1}}^{n} \tilde{w}_{i_1,\ldots,i_d} f(x_{i_1}^{(1)}, \ldots, x_{i_d}^{(d)}), \quad \text{with} \quad \tilde{w}_{i_1,\ldots,i_d} = \prod_{r=1}^{d} w_{i_r}. \tag{11.1.4}$$

Note that in such case, the total number of function evaluations is $M = n^d$

The **curse of the dimensionality** becomes evident once we look into the accuracy of such multi-dimensional quadrature rule. For instance, we know that one-dimensional Simpson's rule is forth order accurate, i.e.

$$I - I_S = \mathcal{O}(h^4). \tag{11.1.5}$$

However, unlike in one-dimensional setting, where interval size $h$ is inversely proportional to the total number of quadrature points $M$ (since for $d = 1$ we have $M = n$). i.e.

$$h = \frac{b_1 - a_1}{n} = \mathcal{O}(n^{-1}) = \mathcal{O}(M^{-1}), \tag{11.1.6}$$

in the multi-dimensional ($d > 1$) case we have an exponential dependence with the exponent $d$,

$$h = \frac{(b_1 - a_1)}{n} = \cdots = \frac{(b_d - a_d)}{n} = \mathcal{O}(n^{-1}) = \mathcal{O}(M^{-1/d}). \tag{11.1.7}$$

Taking this into account, the order of accuracy with respect to the number of function evaluations is no longer 4, but $4/d$,

$$I - I_S = \mathcal{O}(M^{-4/d}). \tag{11.1.8}$$

In general, an order $s$ scheme is order $s/d$ in $d$ dimensions. For large dimensions $d$, the order of accuracy is hence significantly reduced and the required computational cost $M = n^d$ could be unfeasible.

## 11.2 Probability background

Monte Carlo methods are stochastic and therefore can only be analyzed from a statistical viewpoint. Hence, we first review some basic principles from probability theory before describing Monte Carlo integration. One of the most fundamental object in probability theory are **random variables** $X$. For simplicity let us first regard discrete random variables. They can be seen as objects which can take values in some subspace of the natural numbers, i.e. $x \in \Omega \subseteq \mathbb{N}$. For each of the values we assign a probability $P(x) \in [0, 1]$. These probabilities are defined via the property that they sum up to one

$$\sum_i P(x_i) = 1. \tag{11.2.1}$$

A classical example of a discrete random variable is a coin toss. Here the realizations are head H and tail T. To map this to the language introduced above we realize that we can identify head and tail with 0 and 1, thus our space is given by $\{\text{Head}, \text{Tail}\} \equiv \{0, 1\} \equiv \Omega$. Our random variable is the outcome of a toss $X$, where for a fair coin we assume the probability for each of the two states to be the same $P(\text{Head}) = P(\text{Tail}) = 0.5$. From this canonical example we can now form one of the most important discrete probability distributions $\mathbb{P}(X)$, the **Binomial distribution**. It gives the probability of trowing $k$ heads in $n$ tosses, given the probability of tossing a single head $\mathbb{P}(X = \text{Head}) = p$

$$P(k) = \binom{n}{k} p^k (1 - p)^{n-k} \tag{11.2.2}$$

In general we do not only regard discrete spaces, but continuous cases. In order to treat them we have to introduce further objects. From a mathematical point of view this requires us to take a step back from the gained view that we can assign a probability to each element of a set, but rather to intervals on the space.

### 11.2.1 Cumulative distribution and density functions

The **cumulative distribution function** $F_X(x)$, or CDF, of a continuous random variable $X$ (i.e. an object which takes value in a subset of the real numbers $x \in \Omega \subseteq \mathbb{R}$) is the probability $P$ that a value chosen from the variable's distribution is less than or equal to some threshold $x$

$$F_X(x) = P(X \leq x). \tag{11.2.3}$$

The corresponding **probability density function** $p$, or PDF, is the derivative of the CDF:

$$p(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x). \tag{11.2.4}$$

CDFs are always monotonically increasing, which means that the PDF is always non-negative $\mathrm{pdf}(x) \geq 0$. Furthermore they also integrate up to one

$$\int p(x)\mathrm{d}x = 1 \tag{11.2.5}$$

It is important to realize although they share some properties, the PDF can not be identified directly with the probability of an event as it was the case in the discrete setting. The introduces objects give rise to an important relationship and other property of the PDF, namely that the probability that a random variable lies within an interval is given by

$$P(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x. \tag{11.2.6}$$

A typical example here is the **uniform distribution** $\mathcal{U}([a,b])$, which is defined on an interval $[a,b]$ and whose PDF reads

$$p_{\mathcal{U}}(x) = \frac{1}{b-a} \tag{11.2.7}$$

This can be readily be generalized to a domain $\Omega \subseteq \mathbb{R}^n$, where we then find $p_{\mathcal{U}}(\vec{x}) = \frac{1}{|\Omega|}$ with the volume of the domain denoted by $|\Omega|$. The second very important continuous distribution is the **normal distribution** $\mathcal{N}(\mu, \sigma^2)$ which is defined via it's mean $\mu$ and standard deviation $\sigma$. It's pdf reads

$$p_{\mathcal{N}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{11.2.8}$$

### 11.2.2 Expected value and variance

The **expected value** or **expectation** $\mathbb{E}$ of a random variable $X$ over a domain $\Omega$ is defined as

$$\mathbb{E}[X] = \langle X \rangle = \int_\Omega x p(x) \mathrm{d}x. \tag{11.2.9}$$

We remark that given $X$ is a random variable also $f(X)$ is a random variable, thus computing the expectation value of a random variable is closely related to integration. In the discrete setting this reduces to the **mean**

$$\mathbb{E}(x) = \bar{x} = \sum_i x_i P(x_i) \tag{11.2.10}$$

It is easy to see from this definition, that this is a linear operation, i.e. for any constant $a, b$ and random variables $X, Y$ we have

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y] \tag{11.2.11}$$

The **variance** Var of a random variable $X$ over a domain $\Omega$ is defined as

$$\mathrm{Var}[X] = \sigma^2[X] = \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big], \tag{11.2.12}$$

where $\sigma$, the **standard deviation**, is the square root of the variance. Using linearity of the expectation value it is possible to derive a simpler expression for the variance:

$$\sigma^2[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{11.2.13}$$

If two random variables $X, Y$ are **uncorrelated**, i.e. for the expectation value of a product of two variables we find

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y], \tag{11.2.14}$$

then the linearity also holds for the variance

$$\sigma^2\Big[\sum_i a_i Y_i\Big] = \sum_i a_i^2 \sigma^2[Y_i]. \tag{11.2.15}$$

Another property is **independent**, in which case the expectation of products of random variables can be factored, nonetheless this property is stronger and not all uncorrelated variables are independent.

## 11.3 Monte Carlo Integration

Before going into the formal details, let us regard the problem of estimating the value of $\pi$. To achieve this using integration techniques, we recall that the area of a circle with radius $r$ is $\pi r^2$. Via the computation of the area of a quarter of a circle ("the pond") with radius $r = 1$, we would obtain the value $\pi$, see Figure 11.1. This means, that we need to integrate function $f : [0,1]^2 \to \mathbb{R}$, defined by

$$f(x,y) = \begin{cases} 1 & \text{if } \sqrt{x^2+y^2} \le 1, \\ 0 & \text{else,} \end{cases} \tag{11.3.1}$$

over the domain $[0,1]^2$.

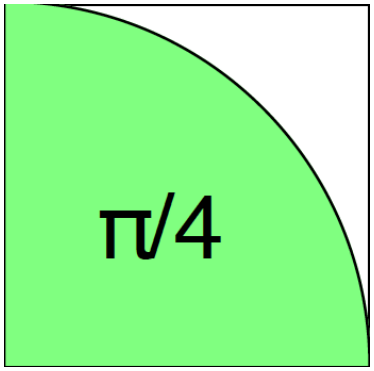


Figure 11.1: The area of the quarter circle with radius $r = 1$ is equal to $\pi/4$.

Instead of using the ideas from the previous lectures we want to find a way to approximate the value of the integral by sampling random points from the domain count how many stones hit the pond (i.e. how many coordinates $(x, y)$ lie inside the circle). Let us make this idea formal:

For a given multi-dimensional integrand $f : \mathbb{R}^d \to \mathbb{R}$, we denote the integral by

$$I = |\Omega|\langle f \rangle, \tag{11.3.2}$$

where we can identify $\langle f \rangle$ as the previously introduced expectation value of the integrand $f$ over $\Omega$, assuming an uniform distribution

$$\langle f \rangle = \frac{1}{|\Omega|} \int_\Omega f(\vec{x}) \mathrm{d}\vec{x}, \qquad |\Omega| = \int_\Omega \mathrm{d}\vec{x}. \tag{11.3.3}$$

To approximate $\langle f \rangle$, instead of evaluating $f$ at $n^d$ locations obtained from a $d$-fold Cartesian product of one-dimensional quadrature rules, we evaluate it at $M$ points $\vec{x}_i$ (called "samples") chosen from an uniform random distribution in $\Omega$ (i.e. regarded as samples from a random variable) and compute the average of all values (perform "sampling")

$$\langle f \rangle \approx \langle f \rangle_M = \frac{1}{M} \sum_{i=1}^{M} f(\vec{x}_i). \tag{11.3.4}$$

Two interpretations of MC estimation are given in Figure 11.2, where MC sampling is performed for a one-dimensional function $f$ on domain $|\Omega| = b - a$ and the number of samples is set to $M = 4$.

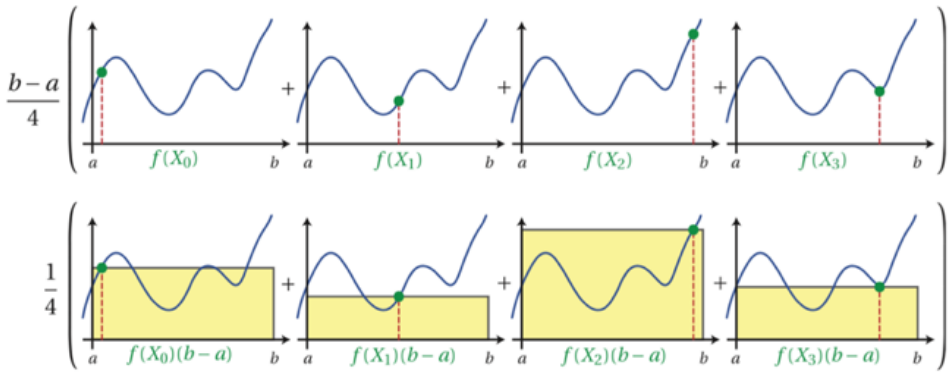


Figure 11.2: Illustration of two interpretations of the Monte Carlo estimator (11.3.4): computing the mean value (or height) of the function and multiplying by the interval length (top), or computing the average of several rectangle rules with random evaluation locations.

Due to the random nature of Monte Carlo sampling (changing random number sequence would result in a slightly different result), the estimate $\langle f \rangle_M$ itself is a random variable. However, assuming that all samples $\vec{x}_i$ are independent, we obtain that the expected value of this estimate $\langle f \rangle_M$ is the exact integral:

$$\mathbb{E}[\langle f \rangle_M] = \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M} f(\vec{x}_i)\right] = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}[f(\vec{x}_i)] = \frac{1}{M}\sum_{i=1}^{M}\langle f \rangle = \frac{1}{M}M\langle f \rangle = \langle f \rangle. \tag{11.3.5}$$

Such estimators are called unbiased ("true") estimators. Furthermore, as we increase the number of samples $M$, the estimator $\langle f \rangle_M$ becomes a close and closer approximation of $\langle f \rangle$. Due to the Strong Law of Large Numbers, in the limit we can guarantee that we have the exact solution:

$$\mathbb{P}\left\{\lim_{M\to\infty}\langle f \rangle_M = \langle f \rangle\right\} = 1. \tag{11.3.6}$$

### 11.3.1 Estimating the error

To compare the accuracy and efficiency of Monte Carlo sampling to other quadrature methods, we need to estimate its error. We define the error to be the standard deviation (i.e., the root mean square (RMS) error) of the difference between random estimate $\langle f \rangle_M$ and exact deterministic integral $\langle f \rangle$,

$$\varepsilon_M = \sqrt{\mathrm{Var}[\langle f \rangle_M - \langle f \rangle]} \tag{11.3.7}$$

Using the expression from equation 11.2.13, the linearity of the expectation value and the fact that $\mathbb{E}[\langle f \rangle_M] = \langle f \rangle$ and is a deterministic quantity, we can further simplify $\varepsilon_M$ to

$$\varepsilon_M = \sqrt{\mathrm{Var}[\langle f \rangle_M]}. \tag{11.3.8}$$

Next, we will derive a closed form expression for the error $\varepsilon_M$ in terms of the variance of the integrand $f$ and the number of samples $M$. Let us therefore plug in the definition of $\langle f \rangle_M$ into our simplified expression for the variance (equation 11.2.13), use the that $\langle \langle f \rangle_M \rangle = \langle f \rangle$ and that $f(\vec{x}_i)$ and $f(\vec{x}_j)$ are independent if $i \neq j$:

$$\begin{aligned}
\varepsilon_M^2 = \mathrm{Var}[\langle f \rangle_M] &= \langle \langle f \rangle_M^2 \rangle - \langle \langle f \rangle_M \rangle^2 \\
&= \frac{1}{M^2} \sum_{i,j=1}^{M} \Big( \mathbb{E}[f(\mathbf{x}_i) f(\mathbf{x}_j)] - \langle f \rangle^2 \Big) \\
&= \frac{1}{M^2} \sum_{i=1}^{M} \Big( \mathbb{E}[f(\mathbf{x}_i)^2] - \langle f \rangle^2 \Big) + \frac{1}{M^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{M} \Big( \underbrace{\mathbb{E}[f(\mathbf{x}_i) f(\mathbf{x}_j)]}_{= \mathbb{E}[f(\mathbf{x}_i)] \mathbb{E}[f(\mathbf{x}_j)] = \langle f \rangle^2} - \langle f \rangle^2 \Big) \\
&= \frac{1}{M^2} \sum_{i=1}^{M} \Big( \langle f^2 \rangle - \langle f \rangle^2 \Big) \\
&= \frac{1}{M^2} M \Big( \langle f^2 \rangle - \langle f \rangle^2 \Big) = \frac{\langle f^2 \rangle - \langle f \rangle^2}{M} = \frac{\mathrm{Var}[f]}{M}.
\end{aligned} \tag{11.3.9}$$

We have obtained that the variance of the Monte Carlo estimator $\langle f \rangle_M$ is $M$ times smaller than the initial variance of the integrand $f$. Hence, Monte Carlo method is $1/2$-order accurate, i.e.

$$\varepsilon_M = \sqrt{\frac{\mathrm{Var}[f]}{M}} = \mathcal{O}(M^{-1/2}). \tag{11.3.10}$$

Notice, that order $1/2$ is lower than any of the quadrature rules discussed so far. Order $1/2$ implies that in order to reduce the error by a factor of 2, we would need to evaluate 4 times as many samples. However, there is a trade-off here: the order does *not* depend on the dimension $d$ of the integrand $f$. Having such dimension-independent accuracy, Monte Carlo sampling has a significant advantage for very high-dimensional integrals. As an example we can compare Monte Carlo sampling to Simpson's rule with order of accuracy $4/d$: we see that Monte Carlo quadrature is more efficient (meaning $1/2 > 4/d$) for any dimension $d$ higher than 8.

**Remark** The error $\varepsilon_M$ of the Monte Carlo estimator is *not* a strict bound of the error, meaning that the following inequality does is *not always* satisfied:

$$|\langle f \rangle - \langle f \rangle_M| \leq \varepsilon_M. \tag{11.3.11}$$

Instead, $\varepsilon_M$ describes the most probable values of the error, i.e. for large number of samples $M$, we expect

$$|\langle f \rangle - \langle f \rangle_M| < \begin{cases} \varepsilon_M, & \text{with probability of 68\%,} \\ 2\varepsilon_M, & \text{with probability of 95\%,} \\ 3\varepsilon_M, & \text{with probability of 99\%.} \end{cases} \tag{11.3.12}$$

### 11.3.2 Summary

Recipe for Monte Carlo Integration of uncorrelated data is as follows:

1. Sample a points $\mathbf{x}_i$ from an uniform distribution and evaluate the integrand $f$ to get random variables $f(\mathbf{x}_i)$.

2. Store the number of samples, the sum of values, and the sum of squares

$$M, \quad \sum_{i=1}^{M} f(\mathbf{x}_i) \quad \sum_{i=1}^{M} f(\mathbf{x}_i)^2. \tag{11.3.13}$$

3. Compute the mean as the estimate of the expectation (normalized integral)

$$\frac{I}{|\Omega|} = \langle f \rangle \approx \langle f \rangle_M = \frac{1}{M} \sum_{i=1}^{M} f(\mathbf{x}_i). \tag{11.3.14}$$

4. Estimate the variance using the unbiased sample variance, which is given as follows

$$\mathrm{Var}[f] \approx \frac{M}{M-1} \left( \frac{1}{M} \sum_{i=1}^{M} f(\mathbf{x}_i)^2 - \langle f \rangle_M^2 \right) = \frac{M}{M-1} \Big( \langle f^2 \rangle_M - \langle f \rangle_M^2 \Big), \tag{11.3.15}$$

   and use it to estimate the error

$$\varepsilon_M = \sqrt{\frac{\mathrm{Var}[f]}{M}} \approx \sqrt{\frac{1}{M-1} \Big( \langle f^2 \rangle_M - \langle f \rangle_M^2 \Big)}. \tag{11.3.16}$$

## 11.4 Non-Uniform Distributions

In practical applications - as to compute macroscopic quantities of a liter of water in the statistical physics context - we usually apply Monte Carlo Integration to estimate the expected value of a function $f(x)$ over some probability density function $p(x)$, i.e.,

$$\mathbb{E}_p[f] = \int f(x) p(x) \, \mathrm{d}x. \tag{11.4.1}$$

The Monte Carlo estimation of this expected value is

$$\mathbb{E}_p[f] \approx \frac{1}{M} \sum_{i=1}^{M} f(x_i) \tag{11.4.2}$$

where $\{x_i | i = 1, \ldots, M\}$ is a set of $M$ samples drawn from probability distribution with probability density function $p(x)$. Recall that for the integrals of the form $I = \int_\Omega f(x) \, \mathrm{d}x$ that we considered so far, we have

$$I = |\Omega| \langle f \rangle = |\Omega| \mathbb{E}_p[f], \tag{11.4.3}$$

where

$$p(x) = \begin{cases} \frac{1}{|\Omega|}, & \text{if } x \in \Omega, \\ 0, & \text{otherwise.} \end{cases} \tag{11.4.4}$$

Hence, we would need to draw samples from uniform distribution. Uniformly distributed random numbers can be obtained using the so-called pseudo-random number generators, available in most of the commercial and free software. Hence, we will assume that such random number are available. For other problems where $p(x)$ is non-uniform, we would like to generate non-uniform random numbers for estimating the expected value $\mathbb{E}_p[f]$.

### 11.4.1 Inverse Transform Sampling

Let X be a random variable (RV) with a probability density function (PDF) $p_X(x)$ and cumulative density function (CDF) $F_X(x)$. Samples from the PDF $p_X(x)$ can be generated using the **Inverse Transform Sampling** method, which makes use of the samples $u^{(i)}$, $i = 1, \ldots, N$ obtained from a uniformly distributed random variable $U \sim \mathcal{U}([0,1])$. In order to do so we want to construct a transformation between the values of the random variables $X$ and $U$

$$x = g(u) \tag{11.4.5}$$

For our uniform random variable $U$, one has that $p_U(u) = 1$ for $u \in [0,1]$ and the cdf correspondingly reads

$$F_U(u) = \int_0^u p_U(s)\mathrm{d}s = u \tag{11.4.6}$$

Using the fact that for any cdf we find $F_X(x) \in [0,1]$ where the value grows monotonically we can define the value $x$ via the requirement that

$$F_X(x) = u \tag{11.4.7}$$

Thus the following transformation is true

$$x = F_X^{-1}(u) \tag{11.4.8}$$

which means that the values of $x$ are given by the inverse of the CDF $F_X(x)$ which also this inverse represents exactly the function $g(u)$. This transformation allows to draw samples $x^{(i)}$, $i = 1, \cdots, N$ from the PDF $p_X(x)$ as

$$x^{(i)} = F_X^{-1}(u^{(i)}) \tag{11.4.9}$$

where $u^{(i)}$ are samples from the uniform distribution over interval $[0,1]$.

Another (more formal way) way to see this and easily check whether a given transformation fulfills the wished relation goes vie the substitution rule. Consider the following integral and the transformation $\vec{x} = g(\vec{u})$

$$\int_{g(U)} p(\vec{x})\mathrm{d}\vec{x} = \int_U p(\vec{u})|J(\vec{x})|^{-1}\mathrm{d}\vec{u} \tag{11.4.10}$$

Where $J$ denotes the Jacobian and $g(U)$ the transformed domain.

**Sampling from the Exponential Distribution** Use the inverse transform sampling method to sample from the exponential distribution

$$p(x) = \lambda e^{-\lambda x}, x > 0 \tag{11.4.11}$$

The CDF of the exponential distribution is

$$F(x) = \int_0^x p(x)dx = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x} \tag{11.4.12}$$

Thus the transformation $x = g(u)$ which is obtained from $x = F^{-1}(u)$ or equivalently $F(x) = u$ is derived by solving

$$1 - e^{-\lambda x} = u \tag{11.4.13}$$

with respect to $x$ to yield

$$x = -\frac{1}{\lambda}\ln(1-u) \tag{11.4.14}$$

The transformation (11.4.14) between the random variable $X$ and the standard uniform variable $U$ defines an exponentially distributed random variable $X$ and it allows to draw samples $x^{(i)}$, $i = 1,\ldots,N$ from the PDF $p(x)$ as

$$x^{(i)} = -\frac{1}{\lambda}\ln(1-u^{(i)}) \tag{11.4.15}$$

where $u^{(i)}$, $i = 1,\ldots,M$ are samples drawn from the standard uniform distribution.

**Sampling from Normal (Gaussian) distribution** The inverse transform sampling method requires the inversion of the CDF $F_X(x)$. This may be time consuming for cases where $F_X^{-1}(u)$ is not known in closed form as, for example, the Normal distribution.

For Normal distribution, an alternative, called **Box-Muller Transformation**, yields an exact method that uses the inverse transform method to convert two independent uniform random variables $u_1, u_2 \sim \mathcal{U}([0,1])$ into two independent normally distributed random variables $x_1, x_2 \sim \mathcal{N}(0,1)$ via the following transformation

$$\begin{aligned} x_1 &= \sqrt{-2\ln(u_1)}\cos(2\pi u_2) \\ x_2 &= \sqrt{-2\ln(u_1)}\sin(2\pi u_2) \end{aligned} \tag{11.4.16}$$

To see that this indeed holds we invert the above equation and find

$$\begin{aligned} u_1 &= \exp\left(-\frac{x_1^2 + x_2^2}{2}\right) \\ u_2 &= \frac{1}{2\pi}\tan^{-1}\left(\frac{x_2}{x_1}\right) \end{aligned} \tag{11.4.17}$$

Calculating the inverse Jacobean of the above transformation and taking the determinant implies the result. Intuitively we can motivate this approach by regarding an integral over the product of two Gaussian probability densities with $\sigma = 1$ and $\mu = 0$ over the unit sphere takes and transforming it to polar coordinates

$$\frac{1}{2\pi} \iint\limits_{x_1^2+x_2^2 \leq r^2} \exp\left(-\frac{x_1^2+x_2^2}{2}\right) \mathrm{d}x_1 \mathrm{d}x_2 = \frac{1}{2\pi} \int_0^{2\pi} \int_0^r \exp\left(-\frac{r^2}{2}\right) r \,\mathrm{d}r \,\mathrm{d}\varphi = 1 - \exp\left(-\frac{r^2}{2}\right) \tag{11.4.18}$$

Setting the right hand side to be $u_1$ and solve for $r$ we find the first factor of equation 11.4.16. The second factor corresponds to the transformation of the uniformly distributed angle on the unit sphere.

In the literature we also find a second form, the so called **Marsaglia polar method**, where the transformation reads

$$\begin{aligned} x_1 &= u_1 \left(\frac{-2\ln(r^2)}{r^2}\right)^{1/2} \\ x_2 &= u_2 \left(\frac{-2\ln(r^2)}{r^2}\right)^{1/2} \end{aligned} \tag{11.4.19}$$

Here we use the complex representation of points on a circle.

## 11.5 Rejection Sampling

Another method for generating random numbers according to a distribution, suggested by von Neumann is Rejection Sampling. In some sense it is a generalization of the ideas we used for the integration in the uniform case. One should find a simple distribution with probability density function $h(x)$ from which we already know how to generate samples, with the property that $h(x)$ bounds $p(x)$, i.e. such that

$$p(x) < \lambda h(x) \tag{11.5.1}$$

for some $\lambda \in \mathbb{R}$. For graphical explanation, refer to Figure 11.3. Then the algorithm continues as follows:

1. draw a random sample $x$ from distribution $h(x)$,
2. draw a uniform random number $u$ in the interval $[0,1]$
3. accept $x$ if $u < \dfrac{p(x)}{\lambda h(x)}$, otherwise reject (forget) $x$,
4. continue the same procedure until sufficiently many (accepted) samples generated.

One of the easiest (not always the best or even appropriate) choice for $h(x)$ is a uniform distribution, as depicted in Figure 11.3. In this case the method is equivalent to integration approach. However,

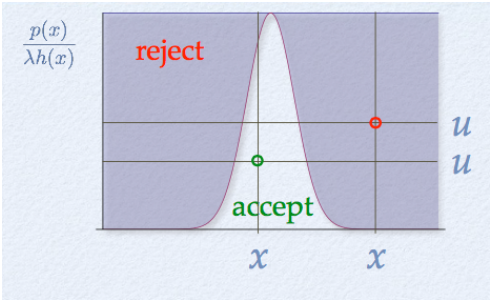


Figure 11.3: Rejection Sampling method

the probability of rejection increases exponentially with $d$, i.e., for high dimensional distributions ($d$ is large), we are facing the same curse of dimensionality that we encountered for other deterministic Quadrature rules.

## 11.6 Importance Sampling

For distribution functions that are highly irregular, multi-modal and peaked, Rejection sampling will waste a lot of effort in regions of low importance (i.e., rejection occurs significantly more often than acceptance). An example of this kind of function can be seen in Figure 11.4 (left). Ideally, we would like to focus our sampling effort in the area where the distribution has more volume in order to explore it better. For example, imagine that you want to sample a rare event. You know, for instance, that you have a distribution function that shows a significant basin around a tiny support of $10^{-5}$ around 0. We would like to bias the random draws so that the majority of the "candidate" points $x$ are being selected around that area; however, such bias needs to be accounted for later. This idea leads to a method called Importance Sampling.

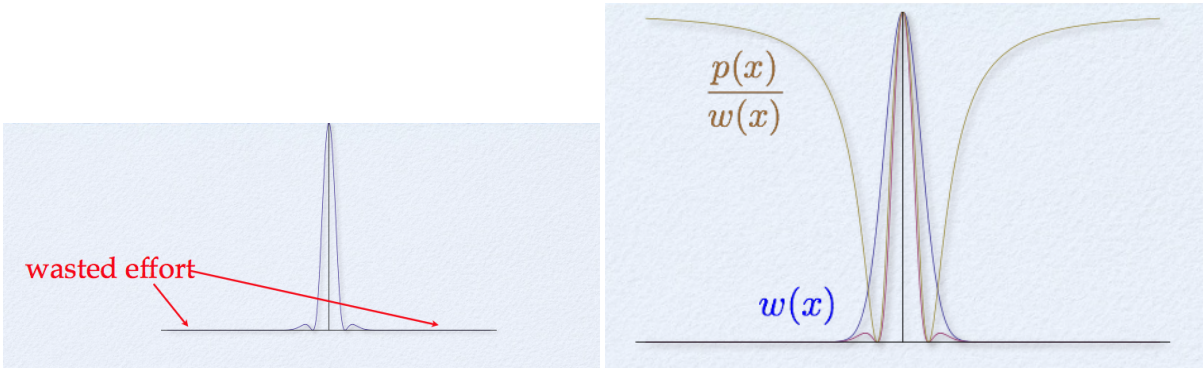


Figure 11.4: Left: Wasted effort while choosing sampling points in areas of low importance. Right: Illustration of usage of importance sampling to overcome difficulties arising from uniform sampling strategies.

Importance sampling allows us to draw samples $x$ that are distributed as probability $w(x)$, instead of a

uniform distribution. We compensate for the bias by normalizing $p(x)$ by the very same "importance" function $w(x)$ and sample $p(x)/w(x)$ instead, resulting in the following integral:

$$\langle f \rangle_p = \int_a^b f(x) \frac{p(x)}{w(x)} w(x) dx \approx \frac{1}{M} \sum_{i=1}^{M} f(x_i) \frac{p(x_i)}{w(x_i)}, \tag{11.6.1}$$

with each $x_i$ being sampled from distribution $w(x)$. For an example distribution $w(x)$, refer to Figure 11.4 (right).

**Exam checklist**

After this class, you should understand the following points regarding numerical integration:

- How can quadrature rules be extended to arbitrary dimensions using the Cartesian product approach.
- What is the curse of dimensionality?
- How does Monte-Carlo integration work? What are its properties?
- How can be generate random numbers from an arbitrary distribution, given uniformly distributed numbers?
- How do Inverse-, Rejection- and Importance-Sampling work?