# MA3227 Numerical Analysis II

## Lecture 22: Monte Carlo Methods

Simon Etter


2019/2020

# Monte Carlo Methods

**Introduction**

Monte Carlo refers to a broad class of algorithms which use randomness to compute a deterministic outcome.

These methods and their name originated from the Manhattan project when the United States developed the first atomic bombs. Since this work was secret, the scientists working on the project required a code name for the methods that they were using. They settled on Monte Carlo in reference to a casino of the same name.

Monte Carlo methods tend to be quite slow, but they are a useful method of last resort when more traditional algorithms break down.

The example on the following slides will help make this clearer.

# Monte Carlo Methods

**Introductory example**

Assume we want to compute the integral

$$I := \int_0^1 \dots \int_0^1 f(y_1, \dots, y_d)\, dy_1 \, \dots \, dy_d.$$

Applying a one-dimensional quadrature rule $(x_k, w_k)_{k=1}^n$ with error $e_n = \mathcal{O}\left(n^{-p}\right)$ repeatedly, we obtain

$$\begin{aligned} I &= \int_0^1 \dots \int_0^1 \left( \sum_{k_1=1}^{n} f(x_{k_1}, y_2, \dots, y_d)\, w_{k_1} + \mathcal{O}\left(n^{-p}\right) \right) dy_2 \, \dots \, dy_d \\ &= \dots \\ &= \sum_{k_d=1}^{n} \dots \sum_{k_1=1}^{n} f(x_{k_1}, \dots, x_{k_d})\, w_{k_1} \, \dots \, w_{k_d} + \mathcal{O}\left(n^{-p}\right). \end{aligned}$$

Observation: we need $N = n^d$ function evaluations (all combinations of $k_1, \dots, k_d \in \{1, \dots, n\}$) to achieve a $\mathcal{O}\left(n^{-p}\right) = \mathcal{O}\left(N^{-p/d}\right)$ error, i.e. the order of convergence in terms of $N$ decreases for increasing $d$.

# Monte Carlo Methods

**Curse of dimensionality**

The observation on the previous slide applies to a vast number of algorithms: the accuracy scales with the "effort per dimension" $n$, but the overall effort scales with $n^d$.

This phenomenon is called the curse of dimensionality and often makes it virtually impossible to do computations in dimensions $d$ larger than about 4 or 5.

Problems with dimensions $d > 5$ are more common than one might think. For example, the probability that someone has the corona virus is a function

$$\underbrace{\mathbb{R}}_{\text{body temperature}} \times \underbrace{\{0,1\}}_{\text{cough}} \times \underbrace{\{0,1\}}_{\text{runny nose}} \times \underbrace{\{0,1\}}_{\text{travel history}} \times \underbrace{\{1,\dots,5\}}_{\text{ethnicity}} \times \underbrace{\{1,\dots,193\}}_{\text{nationality}} \to [0,1].$$

Assuming we need 40 points to cover a reasonable range of body temperatures, this means we need to store

$$40 \times 2 \times 2 \times 2 \times 5 \times 193 = 308'800$$

numbers to represent this function. Moreover, each additional variable in our model multiplies the memory requirements.

# Monte Carlo Methods

**Introductory example (continued)**

The curse of dimensionality can be overcome by reinterpreting $I$ as the expectation value

$$I = \mathbb{E}[f(X_1, \ldots, X_d)], \qquad X_i \overset{\text{iid}}{\sim} \text{Uniform}[0, 1].$$

I will define the precise meaning of $X_i \overset{\text{iid}}{\sim} \text{Uniform}[0, 1]$ later.

Expectations can be computed by taking the average of a sufficiently large number of samples $x_i^{(k)}$ of the $X_i$, i.e.

$$\mathbb{E}[f(X_1, \ldots, X_d)] = \frac{1}{N} \sum_{k=1}^{N} f\left(x_1^{(k)}, \ldots, x_d^{(k)}\right) + \mathcal{O}\left(N^{-1/2}\right).$$

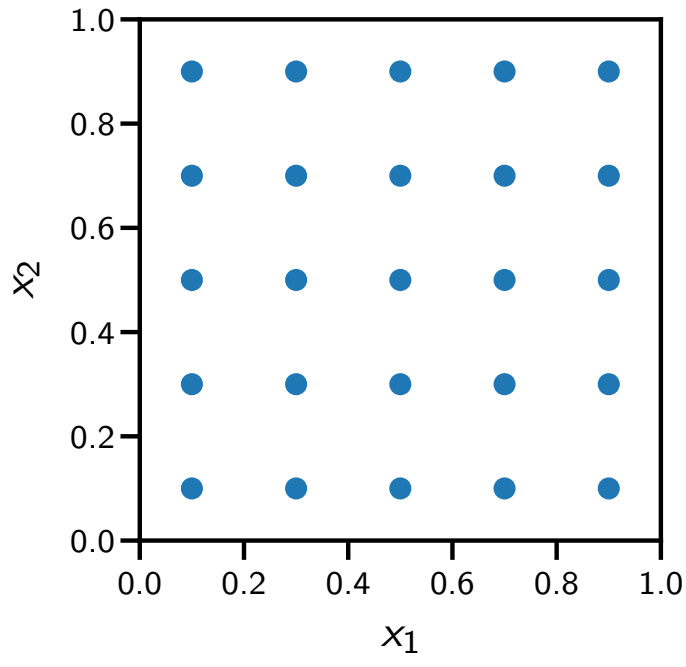
We will discuss the $\mathcal{O}(N^{-1/2})$ error estimate later.

Observation: $N$ evaluations of $f$ lead to $\mathcal{O}\left(N^{-1/2}\right)$ error regardless of $d$. In particular, Monte Carlo is better than midpoint rule ($p = 2$) for $d > 4$.

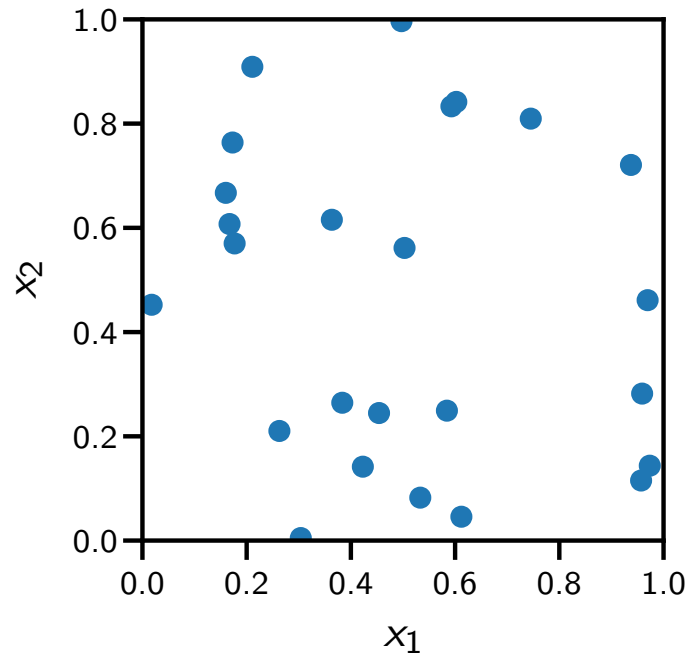
See `convergence()` for numerical demonstration, and see the next slide for a pictorial illustration.

**Introductory example (continued)**

Distribution of the function evaluation points $(x_1, x_2)$.




Midpoint




Monte Carlo

# Monte Carlo Methods

## Monte Carlo methods

In abstract terms, Monte Carlo corresponds to the approximation

$$\mathbb{E}[F] \approx \frac{1}{N}\sum_{k=1}^{N} F_k \qquad \text{where} \qquad F, F_k \overset{\text{iid}}{\sim} \mathcal{F}, \tag{1}$$

i.e. it amounts to estimating expectations by taking the mean of a sufficiently large number of samples as illustrated on the previous slide.

## Outlook

The aim of the following slides is to clarify and explain various technical details in (1) by revisiting some key definitions and results from probability theory.

I strongly recommend that you treat the following simply as an abstract mathematical theory in terms of sets and functions at first. Of course, our aim will be to eventually use this theory to understand why and how the Monte Carlo method works, but that discussion will be easier once we have a firm grasp on the underlying mathematics.

# Monte Carlo Methods

**Disclaimer**

As usual in this module, I will take an engineer's approach to mathematics, i.e. I will ignore many technical details like measurability or $\sigma$-algebras since they are hardly ever relevant in applications.

# Monte Carlo Methods

**Def: Measure on a set $\Omega$**

Function $P$ mapping subsets of $\Omega$ to nonnegative real numbers such that

$$P(\{\}) = 0 \quad \text{and} \quad P(A \cup B) = P(A) + P(B) \text{ if } A, B \subset \Omega \text{ are disjoint.}$$

**Def: Probability measure on a set $\Omega$**

A measure $P$ such that $P(\Omega) = 1$.

**Def: Probability space**

A pair $(\Omega, P)$ where $P$ is a probability measure on $\Omega$

**Def: Random variables**

Any function $X : \Omega \to \Xi$ defined on a probability space $(\Omega, P)$.

**Def: Distribution of a random variable $X : \Omega \to \Xi$**

The probability measure $\hat{P}$ on $\Xi$ given by $\hat{P}(A) = P\big(X^{-1}(A)\big)$.

It is common practice to write $P(X \in A)$ instead of $P(X^{-1}(A))$, and $P(X_1 \in A_1, X_2 \in A_2)$ instead of $P\big(X_1^{-1}(A_1) \cap X_2^{-1}(A_2)\big)$.

This can be confusing if we forget that $P(X \in A)$ is just a shorthand for $P(X^{-1}(A))$, so try your best not to forget this.

# Monte Carlo Methods

## Discussion

It is possible and common practise to talk about random variables and their distributions without ever specifying what the underlying probability space is.

For example, when we say $X \sim \text{Uniform}[0,1]$ ("$X$ is uniformly distributed in $[0,1]$"), we mean that $X : \Omega \to [0,1]$ is a function defined on some unspecified probability space $\Omega$ such that

$$\begin{aligned} P(X \in [a,b]) &= P\big(X^{-1}([a,b])\big) \\ &= P\big(\{\omega \in \Omega \mid X(\omega) \in [a,b]\}\big) \\ &= b - a \end{aligned}$$

for all $a, b$ such that $0 \leq a \leq b \leq 1$.

There are countless ways how we could construct such an $X$. The simplest construction is to choose $\Omega = [0,1]$, $P([a,b]) = b - a$ and $X(\omega) = \omega$, but equally well we could also set $X(\omega) = 1 - \omega$, or $\Omega = [0,2]$, $P([a,b]) = \frac{b-a}{2}$, $X(\omega) = \frac{\omega}{2}$, or an even more complicated construction.

The point is that when we say $X \sim \text{Uniform}[0,1]$, we do not care how $X$ comes about as long as $P(X \in [a,b]) = b - a$.

# Monte Carlo Methods

**Representations of distributions**

Recall that " distribution of a random variable $X : \Omega \to \Xi$ " refers to the probability measure $\hat{P}(A) = P(X^{-1}(A)) = P(X \in A)$ on $\Xi$.

Furthermore, recall that a measure on $\Xi$ is just a function mapping subsets of $\Xi$ to nonnegative real numbers.

There are many ways to represent measures, and correspondingly there are many ways to represent distributions.

The three most common ways to represent distributions are listed on the next slide.

# Monte Carlo Methods

**Representations of distributions (continued)**

- Probability mass function (PMF) of $X : \Omega \to \Xi$ where $\Xi$ is discrete: a function $p : \Xi \to [0, 1]$ such that for all $A \subset \Xi$

$$P(X \in A) = \sum_{x \in A} p(x).$$

- Probability density function (PDF) of $X : \Omega \to \mathbb{R}^n$: a function $f : \mathbb{R}^n \to [0, \infty)$ such that for all $A \subset \mathbb{R}^n$

$$P(X \in A) = \int_A f(x)\, dx.$$

- Cumulative distribution function (CDF) of $X : \Omega \to \mathbb{R}$: a function $F : \mathbb{R} \to [0, 1]$ such that for all $a \in \mathbb{R}$

$$P(X \leq a) = F(a).$$

# Monte Carlo Methods

**Example**

| | Uniform$[0,1]$ | $\mathcal{N}(\mu, \sigma^2)$ |
|---|---|---|
| Name | Uniform distribution on $[0,1]$. | Normal distribution with mean $\mu$ and standard deviation $\sigma$ |
| PDF | $f(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$ | $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ |
| CDF | $F(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \le x \le 1, \\ 1 & \text{if } x > 1. \end{cases}$ | $F(x) = \int_{-\infty}^{x} f(x)\,dx$ (No explicit formula available) |

**Example**

$$X \sim \text{Bernoulli}(p) \qquad \Longleftrightarrow \qquad P(X = 0) = 1 - p, \quad P(X = 1) = p.$$

# Monte Carlo Methods

**Def: Expectation**

The expectation of a random variable $X : \Omega \to \mathbb{R}^n$ is given by

$$\mathbb{E}[X] = \int X(\omega)\, d\omega$$

where $\int d\omega$ denotes the integral with respect to the probability measure $P$ of the underlying probability space.

Integration with respect to a measure may not have been introduced to you yet. If so, ignore the above and study the below special cases instead.

Again, it is possible and common practice to work with expectations without referring to the underlying probability space $(\Omega, P)$.

For example, one can show that if $X : \Omega \to \Xi$ with $\Xi$ discrete or $\Xi = \mathbb{R}^n$ has probability mass / density $p(x)$, then

$$\mathbb{E}[X] = \sum_{x \in \Xi} x\, p(x) \qquad \text{or} \qquad \mathbb{E}[X] = \int_{\mathbb{R}^n} x\, p(x)\, dx,$$

and similarly for any function $f$ on $\Xi$ we have

$$\mathbb{E}[f(X)] = \sum_{x \in \Xi} f(x)\, p(x) \qquad \text{or} \qquad \mathbb{E}[X] = \int_{\mathbb{R}^n} f(x)\, p(x)\, dx.$$

# Monte Carlo Methods

**Def: Variance**

The variance of a random variable $X : \Omega \to \mathbb{R}^n$ is given by

$$\mathrm{Var}[X] = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big].$$

**Useful formulae**

Let $X, Y : \Omega \to \mathbb{R}^n$ be random variables and $a, b \in \mathbb{R}$. Then,

- $\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$.
- $\mathrm{Var}[aX + b] = a^2\,\mathrm{Var}[X]$.
- $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

# Monte Carlo Methods

**Def: Independence**

A collection of random variables $(X_i : \Omega \to \Xi_i)_{i=1}^n$ are called independent if for all $(A_i \subset \Xi_i)_{i=1}^n$ it holds

$$P\big(X_1 \in A_1, \ldots, X_n \in A_n\big) = \prod_{i=1}^{n} P(X_i \in A_i).$$

One can show that $(X_i)_{i=1}^n$ are independent if and only if

- $f(x_1, \ldots, x_n) = f_1(x_1) \ldots f_n(x_n)$ for all $(x_i)_{i=1}^n$, where $f, f_i$ are the PDFs / PMFs of $(X_i)_{i=1}^n$ and $X_i$, respectively.
- $F(x_1, \ldots, x_n) = F_1(x_1) \ldots F_n(x_n)$ for all $(x_i)_{i=1}^n$, where $F, F_i$ are the CDFs of $(X_i)_{i=1}^n$ and $X_i$, respectively.

Moreover, if $(X_i)_{i=1}^n$ are independent, then

- $\mathbb{E}(X_1 \ldots X_n) = \prod_{i=1}^n \mathbb{E}(X_i)$.
- $(f_i(X_i))_{i=1}^n$ are independent for any collection of functions $f_i$.

# Monte Carlo Methods

**Def: Independently and identically distributed**

We say a collection of random variables $X_1, \ldots, X_n$ is identically and independently distributed according to a distribution $\mathcal{D}$, or $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{D}$, if the $X_i$ are independent and each $X_i \sim \mathcal{D}$.

**Def: Monte Carlo estimate of the expectation**

Assume $F$ is a random variable with distribution $\mathcal{F}$. We then set

$$\tilde{\mathbb{E}}_N[F] = \frac{1}{N} \sum_{k=1}^{N} F_k \qquad \text{where } F_k \overset{\text{iid}}{\sim} \mathcal{F}.$$

# Monte Carlo Methods

**Discussion**

If you followed my advice on slide 7, then you think of the above simply as a collection of abstract definitions in terms of sets and functions. In particular, the Monte Carlo estimate $\tilde{\mathbb{E}}_N[F]$ as defined above is just a function $\Omega \to \Xi$ since it is the sum of functions $F_k : \Omega \to \Xi$.

Of course, the goal of a Monte Carlo simulation is not to produce a function $\tilde{\mathbb{E}}_N[F]$ but rather to determine a single element of $\Xi$ which is close to $\mathbb{E}[F]$. This can be achieved by choosing some fixed $\omega \in \Omega$ and evaluating $\tilde{\mathbb{E}}_N[F](\omega)$.

The reason why this procedure works is because the central limit theorem presented on the next slide asserts that $\tilde{\mathbb{E}}_N[F]$ maps almost all $\omega \in \Omega$ to points $\tilde{\mathbb{E}}_N[F](\omega)$ which are close to $\mathbb{E}[F]$.

# Monte Carlo Methods

## Central limit theorem

Assume $F, F_1, \ldots, F_N \overset{\text{iid}}{\sim} \mathcal{F}$. Then,

$$\tilde{\mathbb{E}}_N[F] \quad \xrightarrow{d} \quad \mathcal{N}\Big(\mathbb{E}[F], \tfrac{1}{N}\mathrm{Var}[F]\Big) \qquad \text{for } N \to \infty.$$

Here, $\xrightarrow{d}$ stands for convergence in distribution. This means that the distribution of the random variable on the left becomes increasingly indistinguishable from a normal distribution with mean and standard deviation as indicated on the right.

The above formulation of the central limit theorem is not quite correct because the right-hand side still involves $N$ and is therefore not a valid limit. The technically correct formulation is

$$\sqrt{\tfrac{N}{\mathrm{Var}[F]}}\Big(\tilde{\mathbb{E}}_N[F] - \mathbb{E}[F]\Big) \quad \xrightarrow{d} \quad \mathcal{N}(0,1).$$

In this module, we will mostly ignore the convergence part and simply assume that

$$\frac{1}{N}\sum_{k=1}^{N} F_k \quad \sim \quad \mathcal{N}\Big(\mathbb{E}[F], \tfrac{1}{N}\mathrm{Var}[F]\Big) \qquad \text{if } N \gtrsim 100.$$

# Monte Carlo Methods

## Interpretation of the central limit theorem

Two slides ago, I claimed that the central limit theorem implies that $\tilde{\mathbb{E}}_N[F]$ maps almost all $\omega \in \Omega$ to points $\tilde{\mathbb{E}}_N[F](\omega)$ close to $\mathbb{E}[F]$. There are at least two ways how we can see that this is true.

*Interpretation 1.* The central limit theorem in particular implies that for $N \gtrsim 100$ we have

$$\mathbb{E}\big[\tilde{\mathbb{E}}_N[F]\big] = \mathbb{E}[F] \quad \text{and} \quad \operatorname{Var}\big[\tilde{\mathbb{E}}_N[F]\big] = \tfrac{1}{N} \operatorname{Var}[F].$$

See the remark at the bottom of the previous slide regarding the origin and implication of the assumption $N \gtrsim 100$.

Furthermore, recall that the variance is defined as

$$\operatorname{Var}\big[\tilde{\mathbb{E}}_N[F]\big] = \mathbb{E}\Big[\Big(\tilde{\mathbb{E}}_N[F] - \mathbb{E}\big[\tilde{\mathbb{E}}_N[F]\big]\Big)^2\Big].$$

Inserting the former equations into the latter yields

$$\mathbb{E}\Big[\Big(\tilde{\mathbb{E}}_N[F] - \mathbb{E}[F]\Big)^2\Big] = \tfrac{1}{N} \operatorname{Var}[F],$$

i.e. $\tilde{\mathbb{E}}_N[F]$ has an error $\sqrt{\frac{1}{N} \operatorname{Var}[F]}$ in expectation.

**Interpretation of the central limit theorem (continued)**

*Interpretation 2.* If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\begin{aligned} P(|X-\mu| > \varepsilon\,\sigma) &= \tfrac{1}{\sqrt{2\pi}\,\sigma} \int_{\mu-\varepsilon\,\sigma}^{\mu+\varepsilon\,\sigma} \exp\Big(-\tfrac{(x-\mu)^2}{2\sigma^2}\Big)\,dx \\ &= \tfrac{1}{\sqrt{2\pi}} \int_{-\varepsilon}^{\varepsilon} \exp\big(-\tfrac{y^2}{2}\big)\,dy \\ &= g(\varepsilon) \end{aligned}$$

where the second line follows by substituting $y = \frac{x-\mu}{\sigma}$, $dy = \frac{dx}{\sigma}$. Moreover, it can be shown that $g(\varepsilon) \to 1$ very quickly for $\varepsilon \to \infty$.

Since the central limit theorem asserts $\tilde{\mathbb{E}}_N[F] \sim \mathcal{N}\big(\mathbb{E}[F], \frac{1}{N}\mathrm{Var}[F]\big)$, we conclude that it is very unlikely that the Monte Carlo estimate $\tilde{\mathbb{E}}_N[F]$ assumes values further than three or four standard deviations $\sqrt{\frac{1}{N}\mathrm{Var}[F]}$ from the exact value $\mathbb{E}[F]$.

**Discussion**

Both of the above interpretations of the central limit theorem lead us to the conclusion that $\tilde{\mathbb{E}}_N[F]$ acts like a funnel: it takes in most (but not necessarily all, see below) of $\Omega$ and maps it into a region around $\mathbb{E}[F]$ with a diameter of roughly $\frac{1}{N}\text{Var}[F]$.

However, the central limit theorem tell us nothing about the worst-case behaviour. It is possible that there are $\omega$ such that $\tilde{\mathbb{E}}_N[F](\omega)$ does not converge to $\mathbb{E}[F]$ regardless of how many samples $N$ that we take.

In a sense, Monte Carlo methods are therefore indeed like gambling: the success or failure of our computations is entirely dependent on whether or not we happen to choose a "good" $\omega$.

**Discussion (continued)**

However, unlike real gambling, Monte Carlo methods allow us to stack the odds heavily in our favour. For example, if we take $N = 10'000$ samples of a random variable $F$ with $\mathrm{Var}[F] = \sigma^2$, then

$$P\big(\big|\tilde{\mathbb{E}}_N[F] - \mathbb{E}[F]\big| > 0.1\,\sigma\big) = P\Big(\big|\tilde{\mathbb{E}}_N[F] - \mathbb{E}[F]\big| > 0.1\,\sqrt{N}\,\big(\tfrac{1}{\sqrt{N}}\sigma\big)\Big)$$
$$= 1 - g(0.1\,\sqrt{N}) \approx 10^{-45}.$$

See slide 21 regarding $g(x)$.

Such odds for picking a bad $\omega$ are negligible compared to those of other sources of error (e.g. faulty hardware, programming errors, getting struck by lightning, etc.).

This concludes our discussion of the probability theory underlying Monte Carlo algorithms. The next lecture will show you how to translate this theory into actual code.

# Monte Carlo Methods

**Summary**

- Monte Carlo idea: estimate expectations by taking the mean of a large number of samples.
- Basic probability theory: probability space, random variables, distributions, expectation, variance, independence.
- Central limit theorem: if $F, F_1, \ldots, F_N \overset{\text{iid}}{\sim} \mathcal{F}$, then

$$\frac{1}{N}\sum_{k=1}^{N} F_k \quad \xrightarrow{d} \quad \mathcal{N}\Big(\mathbb{E}[F], \tfrac{1}{N}\mathrm{Var}[F]\Big) \qquad \text{for } N \to \infty.$$