The Monte Carlo method in a nutshell

Ulrik Skre Fjordholm, Nils Henrik Risebro and Håkon Hoel

November 21, 2023

Let us assume you wish to approximate the integral

$$I(f) = \int_0^1 f(x)\,dx$$

for some continuous function $f\colon [0,1] \to \mathbb{R}$. The integral $I(f)$ can be thought of as the average value of f over $[0,1]$, so what we are asking for is: What is the average, or "expected", value of the function f over the interval $[0,1]$? In this note we review the Monte Carlo approximation of $I(f)$, and we prove an error estimate. The impatient reader may skip to the rigorous explanation on pp. 2–10.

If you need to integrate over an arbitrary interval $[a,b] \subset \mathbb{R}$, perform a change of variables.

The Monte Carlo method was invented and developed by Stanislaw Ulam[1] and his colleagues. While being hospitalized for a longer period of time, Ulam spent his time playing solitaire, and became interested in computing the probability p of a game of solitaire coming out successfully. Even for a first-rate mathematician like Ulam, this proved too difficult to compute, but he came up with an approximation that could easily be carried out: Play a large number (say, $M \in \mathbb{N}$) of games of solitaire, and note the number of times n that the game comes out successfully. The true answer will then be approximately $p \approx n/M$.

[1] Stanislaw Ulam (1909–1984) was a Polish–American physicist and mathematician who worked on the Manhattan project to develop the first atomic bomb. Among his many achievements, he is perhaps best known for the Ulam–Teller design of the hydrogen bomb.

Ulam quickly understood that the approach could be used for problems in nuclear physics, and involved his colleague John von Neumann in the endeavour to apply this new *Monte Carlo method*[2] using the newly developed electronic computers. Generally speaking, the Monte Carlo consists of computing a large number of *samples*, and then averaging over these.

[2] Supposedly named after the Monte Carlo Casino in Monaco, which Ulam's uncle frequented.

There are at least three ingredients that make n/M a good approximation of the true probability p:

(i) M must be moderately large,

(ii) the deck must be well-shuffled at the start of each game,

(iii) each game must be independent of the others.

As we will see, the error in the approximation scales as $1/\sqrt{M}$, which explains the first point. For the second point, if the deck isn't well-shuffled then certain starting decks (and hence, certain outcomes of the game) will occur more often than others, making the computation skewed. For the third point, if one experiment influences successive experiments, then again the computation will be skewed.

A quick explanation

Fix some $M \in \mathbb{R}$. The Monte Carlo approximation to $I(f)$ takes M random numbers $X_1, \dots, X_M$, uniformly distributed in the interval

$[0,1]$, and returns

$$I_M(f) = \frac{1}{M}\sum_{m=1}^{M} f(X_m).$$

On a computer you would compute *pseudo-random* numbers, say, using the Python function `random.random()`.

We can experimentally observe that $I_M(f) \to I(f)$ as $M \to \infty$.

Exercise. Implement the Monte Carlo method in your favourite programming language. Test it using $f(x) = \cos(x)$ and $f(x) = e^{-x^2}$, and with $M = 2, 4, 8, \ldots, 2^{10}$.

Independence

Without further assumptions on the random variables $X_1, \ldots, X_M$, we might as well have chosen them to be equal: $X_1 = \cdots = X_M$, reducing the approximation to $I_M(f) = f(X_1)$. As you might imagine, this would be a terrible approximation of $I(f)$. What is lacking is not randomness, but *independency*: The random variables $X_1, \ldots, X_M$ must be independent from one another. Informally speaking, X_m is independent from X_n if knowing the value of $X_m(\omega)$ will not make it any easier to guess the value of X_n.

Successive calls to the Python function `random.random()` (or an analogous function in other programming languages) will generate independent random variables $X_1, X_2, X_3, \ldots$.

A (mostly) rigorous explanation

For completeness, we first present probability spaces in a general notation and setting and then we narrow our scope to the setting that is relevant for this course: so-called continuous random variables.

General random variables

A triplet $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of the sample space Ω, the event space $\mathcal{F}$ and a probability measure $\mathbb{P}$ is called a probability space.

- The samples/outcomes $\omega \in \Omega$ will not necessarily be numbers, vectors or functions; instead, we will treat Ω as an abstract set of objects.
- An event $E \subset \Omega$ is a union of outcomes from Ω, and the event space $\mathcal{F}$ is the set of all events that we need to measure the probability of.
- The probability measure $\mathbb{P} : \mathcal{F} \to [0,1]$ is a mapping from the set of events $\mathcal{F}$ to $[0,1]$. The probability of an event $E \in \mathcal{F}$ is defined as the value $\mathbb{P}(E)$. No event has a negative probability, and no event has a probability larger than 1. In particular,

$$\underbrace{\mathbb{P}(\emptyset)}_{\text{probability of empty set/event}} = 0, \qquad \underbrace{\mathbb{P}(\Omega)}_{\text{probability of full set/event}} = 1,$$

and

$$\mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A) \quad \text{for any} \quad A \in \mathcal{F}.$$

- A (real-valued) *random variable* on a given space $(\Omega, \mathcal{F}, \mathbb{P})$ is a measurable mapping $X : \Omega \to \mathbb{R}^d$ satisfying that for all for all (Borel) sets $B \subset \mathbb{R}^d$, the event

$$\{X \in B\} = \{\omega \in \Omega \mid X(\omega) \in B\} = X^{-1}(B)$$

belongs to $\mathcal{F}$ so that the probability $\mathbb{P}(X \in B)$ is well defined. We note that Ω is the domain of the random variable, and $\mathbb{R}^d$ is referred to as its state space.

A Borel set in $\mathbb{R}^d$ is a set that can be formed by countable unions and intersections of open sets in $\mathbb{R}^d$ and relative complements.

Continuous random variables

A random variable $X : \Omega \to \mathbb{R}^d$ is called a continuous random variable if there exists a non-negative integrable function $\rho : \mathbb{R}^d \to [0, \infty)$ such that

$$\mathbb{P}(X \in B) = \int_B \rho(x)\, dx \qquad \forall B \subset \mathbb{R}^d\,. \tag{1}$$

The function ρ is called the *probability density function* (PDF) X.

Exercise. Constraint: Take $B = \mathbb{R}^d$ in (1) to show that all PDFs must satisfy the following constraint:

$$\int_{\mathbb{R}^d} \rho(x) dx = 1.$$

Any continuous random variable is uniquely described in terms of its PDF. This relationship is often written $X \sim \rho$, meaning X has the density function ρ. Occasionally, one rather writes ρ_X to compactly express that ρ_X is the PDF of X.

Further properties:

- The *expected value* or mean of a random variable $X : \Omega \to \mathbb{R}^d$ is defined by the abstract integral

$$\mathbb{E}[X] = \int_\Omega X(\omega) \mathbb{P}(d\omega).$$

For continuous random variables, it is equivalently defined by

$$\mathbb{E}[X] = \int_{\mathbb{R}^d} x \rho_X(x) dx.$$

The number $\mathbb{E}[X] \in \mathbb{R}^d$ called the *expected value of* X.

- By the rules of integration, one can show that

  $\mathbb{E}$ *has unit mass:* If X is constant, say, $X(\omega) = a$ for all $\omega \in \Omega$ for some $a \in \mathbb{R}$, then $\mathbb{E}[X] = a$.

  $\mathbb{E}$ *is positivity preserving:* If $X \colon \Omega \to [0, \infty)$ is a *nonnegative* random variable, then $\mathbb{E}[X] \geqslant 0$.

  If $f : \mathbb{R}^d \to \mathbb{R}$ is a measurable mapping and $X : \Omega \to \mathbb{R}^d$ is a continuous random variable, then $f(X)$ is also a random variable and

$$\mathbb{E}[f(X)] = \int_{\mathbb{R}^d} f(x) \rho_X(x) dx.$$

  $\mathbb{E}$ *is linear:* By the properties of joint densities (see below), one can show that $\mathbb{E}[\alpha X + Y] = \alpha \mathbb{E}[X] + \mathbb{E}[Y]$ for every $\alpha \in \mathbb{R}$ and continuous random variables $X, Y \colon \Omega \to \mathbb{R}^d$ (and it also holds more generally).

Multiple/multivariate random varialbes

Given two continuous random variables $X, Y\colon \Omega \to \mathbb{R}^d$ with densities $X \sim \rho_X$ and $Y \sim \rho_X$, the tuple (X, Y) is also a continuous density random variable with $\rho_{X,Y}(x, y)$, satisfying that

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \rho_{X,Y}(x, y)dx\, dy = 1$$

and the marginal density properties

$$\rho_X(x) = \int_{\mathbb{R}^d} \rho_{X,Y}(x, y)dy, \quad \rho_Y(y) = \int_{\mathbb{R}^d} \rho_{X,Y}(x, y)dx.$$

Further properties:

- Two random variables are *independent* if for all (Borel) sets in $A, B \subset \mathbb{R}^d$ we have that

  $$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B),$$

  or equivalently (for continuous RV) if their joint density is mulitiplicatively separable

  $$\rho_{X,Y}(x, y) = \rho_X(x)\rho_Y(y)$$

  Informally speaking, independency means that the knowledge of $X(\omega)$ will not enable you to guess what $Y(\omega)$ is, or vice versa.

- Two continous random variables are *identically distributed* if for all (Borel) sets $A \subset \mathbb{R}^d$ we have that

  $$\mathbb{P}(X \in A) = \mathbb{P}(Y \in A),$$

  or equivalently if their densities are equal: $\rho_X = \rho_Y$.

  Identically distributed random variables "visit the same function values equally often", or "have the same likelihood of returning a given value".

  Exercise. Show that if X and Y are identically distributed continuous random variables, then $\mathbb{E}[X] = \mathbb{E}[Y]$.

  And that if X and Y are independent continuous random variables, then

  $$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

- A collection of random variables $X_1, \ldots, X_m$ is called mutually independent if for all (Borel) sets in $A_1, \ldots, A_m \subset \mathbb{R}^d$ we have that

  $$\mathbb{P}(X_1 \in A, X_2 \in A_2, \ldots, X_m \in A_m) = \prod_{i=1}^{m} \mathbb{P}(X_i \in A_i)$$

  or equivalently if their joint density is multiplicatively separable

  $$\rho_{X_1,X_2,\ldots,X_m}(x_1, \ldots, x_m) = \prod_{i=1}^{m} \rho_{X_i}(x_i).$$

- And for any collection of random variables $X_1, \ldots, X_m$, linearity yields that

  $$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} X_i\right] = \sum_{i=1}^{m} \frac{\mathbb{E}[X_i]}{m}$$

Example. A random variable $X\colon \Omega \to \mathbb{R}^d$ is *uniformly distributed* on $[a,b]^d$ if its PDF is given by

$$\rho(x) = \frac{1}{(b-a)^d}\mathbb{1}_{[a,b]^d}(x) = \begin{cases} \frac{1}{(b-a)^d} & x \in [a,b]^d \\ 0 & \text{else.} \end{cases}$$

Random variables with this distribution is often written $X \sim U([a,b]^d)$, which is equivalent to $X \sim (b-a)^{-d}\mathbb{1}_{[a,b]^d}$.

Example. For the random variables $X, Y \sim U([0,1])$ are independent if and only if their joint density is on the form

$$\rho_{X,Y}(x,y) = \mathbb{1}_{[0,1]^2}(x,y)$$

If X and Y are independent, their joint density tells us that we can view (X,Y) as a vector-valued random variable with

$$(X,Y) \sim U([0,1])^2.$$

Exercise. Show that the Gaussian random variable

$$(X,Y) \sim \rho_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\nu^2}} \exp\left(-\frac{x^2+y^2-2\nu xy}{2(1-\nu^2)}\right)$$

with a given (deterministic) constant $\nu \in (-1,1)$, the scalar components X and Y are independent random variables if and only if $\nu = 0$.

The Monte Carlo approximation

We can now define the Monte Carlo method. Assume that we wish to approximate $\mathbb{E}[X]$ for some random variable $X : \Omega \to \mathbb{R}$. Let $X_1, X_2, \ldots, X_M$ be random variables which are mutually independent and which all have the same density ρ_X.

It will soon become clear what $\mathbb{E}[f(X)]$ has to do with $I(f)$.

The *Monte Carlo approximation* to $\mathbb{E}[X]$ is simply a fancy name for the *average* of M of these variables

$$I_M(\omega) = \frac{1}{M}\sum_{m=1}^{M} X_m(\omega) \qquad \forall\, \omega \in \Omega.$$

We note that I_M, being a scaled sum of random variables, is itself a random variable, and that

$$I_M(\omega) \approx \int_{\mathbb{R}^d} x\rho_X(x)dx = \mathbb{E}[X].$$

An error estimate

We now prove an estimate of the error in the Monte Carlo approximation. Since I_M is itself a random variable, we cannot guarantee that the error $\mathbb{E}[X] - I_M(\omega)$ will be small regardless of the seed ω. We will only be able to guarantee that the *average* error is small.

Our measure of error will be the root mean square error

The mean square error measures how much, on average, $I_M(f)$ deviates from $\mathbb{E}[X]$.

$$\mathcal{E}_M = \sqrt{\mathbb{E}\left[\left(\mathbb{E}[X] - I_M\right)^2\right]}.$$

We square both sides and compute:

$$\begin{aligned}\mathcal{E}_M^2 &= \mathbb{E}\left[\mathbb{E}[X]^2 - 2\mathbb{E}[X]I_M + I_M^2\right] \\ &= \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[I_M] + \mathbb{E}\left[I_M^2\right]. \end{aligned}$$

(since $\mathbb{E}$ is linear and has unit mass)

For the second term we can compute

$$\begin{aligned}\mathbb{E}[I_M] &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}[X_m] && \text{(since } \mathbb{E} \text{ is linear)} \\ &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}[X] && \text{(since } X \text{ and } X_m \text{ are identically distributed)} \\ &= \mathbb{E}[X].\end{aligned}$$

For the third term we get

$$\begin{aligned}\mathbb{E}\left[I_M^2\right] &= \mathbb{E}\left[\left(\tfrac{1}{M}\textstyle\sum_{m=1}^{M} X_m\right)\left(\tfrac{1}{M}\textstyle\sum_{n=1}^{M} X_n\right)\right] \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{n=1}^{M}\mathbb{E}\left[X_m X_n\right] && \text{(since } \mathbb{E} \text{ is linear)} \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{\substack{n=1\\n\neq m}}^{M}\mathbb{E}[X_m]\mathbb{E}[X_n] + \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[X_m^2\right] && \text{(since } X_m \text{ and } X_n \text{ are independent when } n \neq m) \\ &= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{\substack{n=1\\n\neq m}}^{M}\mathbb{E}[X]^2 + \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\left[X^2\right] && \text{(since } X_m, X_n \text{ and } X \text{ are identically distributed)} \\ &= \frac{M^2 - M}{M^2}\mathbb{E}[X]^2 + \frac{1}{M}\mathbb{E}\left[X^2\right] \\ &= \left(1 - \frac{1}{M}\right)\mathbb{E}[X]^2 + \frac{1}{M}\mathbb{E}\left[X^2\right].\end{aligned}$$

Inserting these two computations in the expression for the error $\mathcal{E}_M^2$, we get

$$\begin{aligned}\mathcal{E}_M^2 &= \mathbb{E}[X]^2 - 2\mathbb{E}[X]^2 + \left(1 - \frac{1}{M}\right)\mathbb{E}[X]^2 + \frac{1}{M}\mathbb{E}\left[X^2\right] \\ &= \frac{\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2}{M}.\end{aligned}$$

The expression $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$ is the *variance* of the random variable X. Taking square roots we conclude:

Exercise: Show that $\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$.

Theorem 1. *The root mean square error of the Monte Carlo approximation is*

$$\mathcal{E}_M = \frac{\sigma[X]}{\sqrt{M}}$$

where $\sigma[X] = \sqrt{\mathrm{Var}[X]}$.

$\sigma[Y]$ is the *standard deviation* of Y. Both the variance and the standard deviation give an indication of how much, on average, the random variable deviates from its expected value.

Since the standard deviation $\sigma[X]$ is a constant, the Monte Carlo error scales as $M^{-1/2}$. In order to reduce the expected error by a factor $1/2$, you need to increase M by a factor $2^2 = 4$.

Remark 2. Note that we can use the positivity preserving property of expectation to overestimate the variation of X if X is bounded.

In this case, choose constants a and b such that $b \geq \sup_\Omega X$ and $a \leq \inf_\Omega X$. Set $c = (a+b)/2$ and calculate

$-(b-a)/2 \leq X - c \leq (b-a)/2$

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[((X-c) - (\mathbb{E}[X]-c))^2\right] \\
&= \mathbb{E}\left[(X-c)^2\right] - (\mathbb{E}\left[X-c\right])^2 \\
&\leq \mathbb{E}\left[(X-c)^2\right] \\
&\leq \left(\frac{b-a}{2}\right)^2.
\end{aligned}$$

This error estimate says that "in the (root mean square) average sense" you can approximate $\mathbb{E}[X]$ by I_M making a $\mathcal{O}(M^{-1/2})$ error. However, I_M is a random variable, so it would be convenient to have results stating something like:

With 90% probability, $|I_M - \mathbb{E}[X]|$ *is less than* 10^{-3}.

In order to make such statements, we must first define *probability*. For a continuous random variable $X : \Omega \to \mathbb{R}$, we recall that the probability of an event $X \in A$ for any Borel set $A \subset \mathbb{R}$ can be computed as

$$\mathbb{P}(X \in A) = \int_A \rho_X(x)dx.$$

Yes, it's the guy with the polynomials.

Lemma 3 (Chebyshev's inequalities)**.** *Let* $X : \Omega \to \mathbb{R}$ *be a random variable. Inequality 1, if* $\mathbb{E}[|X|] < \infty$*, then for any* $t > 0$

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[|X|]}{t}.$$

Inequality 2, if $\mathrm{Var}[X] < \infty$*, then for any* $t > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathrm{Var}[X]}{t^2}.$$

We restrict ourselves to proving the result for continuous random variables although it is true in general.

Proof. Let $X \sim \rho$. Then, noting that $x/t \geq 1$ for all $x \geq t$, we have that

$$\begin{aligned}
\mathbb{P}(X \geq t) &= \int_{-\infty}^{\infty} \mathbb{1}_{[t,\infty)}(x)\rho(x)dx \leq \int_t^{\infty} \frac{x}{t}\rho(x)dx \\
&\leq \int_{-\infty}^{\infty} \frac{|x|}{t}\rho(x)dx = \mathbb{E}\left[\frac{|X|}{t}\right].
\end{aligned}$$

The last equality used that $f(X) = |X|/t$ is a random variable, and thus that

$$\mathbb{E}[f(X)] = \int_{-\infty}^{\infty} f(x)\rho(x)dx.$$

This proves inequality 1. Inequality 2 follows by a similar argument. Let $A = \{y \in \mathbb{R} \mid |y - \mathbb{E}[X]|^2 \geq t^2\}$, and note that

$$x \in A \implies \frac{|x - \mathbb{E}[X]|^2}{t^2} \geq \mathbb{1}_A(x) \geq 1, \text{and}$$

$$x \notin A \implies \frac{|x - \mathbb{E}[X]|^2}{t^2} \geq \mathbb{1}_A(x) = 0.$$

Hence,

$$\begin{aligned}\mathbb{P}(|X - \mathbb{E}[X]| \geq t) = \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq t^2) &= \int_{-\infty}^{\infty} \mathbb{1}_A(x)\rho(x)dx \\ &\leq \int_{-\infty}^{\infty} \frac{|x - \mathbb{E}[X]|^2}{t^2}\rho(x)dx = \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]}{t^2} = \frac{\mathrm{Var}[X]}{t^2}.\end{aligned}$$

□

Now we can get the kind of probabilistic error bounds that we wanted. Using Chebyshev's second inequality for the random variable $I_M - \mathbb{E}[X]$ we get

$$\begin{aligned}\mathbb{P}(|I_M - \mathbb{E}[X]| \geq \varepsilon) &\leq \frac{1}{\varepsilon^2}\,\mathrm{Var}[I_M - \mathbb{E}[X]] \\ &= \frac{1}{\varepsilon^2}\mathcal{E}_M^2 \leq \frac{\mathrm{Var}[X]}{\varepsilon^2 M} \leq \frac{(\sup_\Omega X - \inf_\Omega X)^2}{4\varepsilon^2 M}, \qquad (2)\end{aligned}$$

where the last inequality holds only if X is bounded.

We obtain that I_M converges in probability to $\mathbb{E}[X]$, which is defined as follows: for any $\varepsilon > 0$,

$$\lim_{M\to\infty} \mathbb{P}(|I_M - \mathbb{E}[X]| \geq \varepsilon) = 0.$$

We note that is a quite weak form of convergence as $|I_M(\omega) - \mathbb{E}[X]| \to 0$ need not hold for any $\omega \in \Omega$ under convergence in probability.

By a more technical argument, one can also show $\mathbb{P}$-almost sure convergence, defined as:

$$\mathbb{P}(\lim_{M\to\infty} I_M \neq \mathbb{E}[X]) = 0.$$

This is a stronger convergence, as it implies that for almost all samples $\omega \in \Omega$

$$|I_M(\omega) - \mathbb{E}[X]| \to 0 \quad \text{as} \quad M \to \infty.$$

Exercise. (Technical and not curriculum, but for those that are curious on the difference between convergence in probability and almost sure convergence.) Let $X_n \sim U([0,1])$ for $n = 1, 2, \ldots$ be a sequence of mutually independent random variables, and let $Y_n := \mathbb{1}_{[0,n^{-1}]}(X_n)$. Show that then $\mathbb{P}(Y_n = 0) = 1 - n^{-1}$ and $\mathbb{P}(Y_n = 1) = n^{-1}$ and that as $n \to \infty$, then Y_n converges in probability to 0, while Y_n does not converge almost surely to 0.

Hint: For the last point, you can use that

$$\mathbb{P}(\lim_{M\to\infty} Y_n \neq 0) = \mathbb{P}(\cap_{n=1}^{\infty} \cup_{m=n}^{\infty} \{Y_n = 1\})$$

and that apply the second Borel-Cantelli lemma.

Application to numerical integration

We choose now a random variable $X \sim U([0,1]^d)$ and consider a measurable mapping $f : \mathbb{R}^d \to \mathbb{R}$. Then, for a sequence of mutually independent random variables $X_i \sim U([0,1]^d)$, the corresponding random variables $Y_i = f(X_i)$ become independent scalar-valued random variables that are identically distributed and

$$\mathbb{E}[f(X)] = \int_{[0,1]^d} f(x)\rho_X(x)\,dx = \int_{\mathbb{R}^d} f(x)\mathbb{1}_{x\in[0,1]^d}\,dx$$

If for example for, $f(x) = \mathbb{1}_{x\in[a,b]^d}$ for some $[a,b]^d \subseteq [0,1]^d$, then we get

$$\mathbb{P}\left(X \in [a,b]^d\right) = \mathbb{E}\left[\mathbb{1}_{[a,b]^d}\right] = \int_{[0,1]^d} \mathbb{1}_{[a,b]^d}(x)\,dx = (b-a)^d,$$

which confirms the intuitive idea of a "uniform distribution". And the Monte Carlo approximation

$$I_M(f) = \frac{1}{M}\sum_{m=1}^{M} f(X_m)$$

will give an approximation of the integral $I(f) = \int_{[0,1]^d} f(x)\,dx$, and the error in this approximation scales as $M^{-1/2}$.

Summing up the two main results in the context of Monte Carlo integration:

$$\mathcal{E}_M(f) \leq \frac{\left(\max_{x\in[0,1]^d}\{f(x)\} - \min_{x\in[0,1]^d}\{f(x)\}\right)}{2\sqrt{M}},$$

and

$$\mathbb{P}\left(\left|I_M(f) - \int_{[0,1]^d} f(\boldsymbol{x})\,d\boldsymbol{x}\right| \geq \varepsilon\right) \leq \frac{\left(\max_{\boldsymbol{x}\in[0,1]^d}\{f(\boldsymbol{x})\} - \min_{\boldsymbol{x}\in[0,1]^d}\{f(\boldsymbol{x})\}\right)^2}{4\varepsilon^2 M}.$$

Note carefully that the root mean square error scales as $M^{-1/2}$, regardless of the dimension d. This is in stark contrast to more standard quadrature methods, whose error scales as $M^{-k/d}$, where k is the accuracy of the quadrature method and M is the number of quadrature points. If d is very large, then the error will converge to zero very slowly – this is the *curse of dimensionality.*

Unlike classic quadrature methods, the Monte Carlo method does not suffer from the curse of dimensionality and its order of convergence is not sensitive to the regularity of the integrand.

Example. For the midpoint method using stepsize $h = M^{-1/d}$ and assuming that $M^{1/d}$ is an integer, the integral of f is approximated

by

$$I_M^{MP}(f) = \sum_{i_1=1}^{M^{1/d}} \cdots \sum_{i_d=1}^{M^{1/d}} f\Big(i_1(h+1/2), i_2(h+1/2), \ldots, i_d(h+1/2)\Big) h^d \,.$$

One can show that if $f \in C^2([0,1]^d)$, then

$$\left| I_M^{MP}(f) - \int_{[0,1]^d} f(x) dx \right| \leq Ch^2 = \mathcal{O}(M^{-2/d}).$$

While measured for instance in root mean square error, the Monte Carlo arguably converges with a higher order than the midpoint method whenever $d > 4$, as it holds that

$$\sqrt{\mathbb{E}\left[\left| I_M(f) - \int_{[0,1]^d} f(x) dx \right|^2 \right]} = \mathcal{O}(M^{-1/2}).$$

The price one pays for lifting the curse of dimensionality is that one only gets probabilistic error estimates and these can also be expensive. Using the error estimate (2), if we wish to be 99% sure that we have an error $\left| I_M(f) - \int_{[0,1]^d} f(x)\, dx \right|$ less than 0.005 (correct to two decimal places), we can choose the number of samples M such that

$$\frac{(\max\{f\} - \min\{f\})^2}{4(0.005)^2 M} \leq 0.01,$$

that is

$$M \geq 1\,000\,000\ (\max\{f\} - \min\{f\})^2 \,.$$