Introduction to Computational Physics

Monte Carlo Methods - I

Andreas Adelmann

Paul Scherrer Institut, Villigen
E-mail: andreas.adelmann@psi.ch

https://moodle-app2.let.ethz.ch/course/view.php?id=15323

Content I

7.1 Introductory Comments I

- ...
- G. Comte de Buffon, *Supplément á l'Historie Naturelle*, vol. 4, 1777. Randomly thrown needles on a slat floor to get π.
- R. Brown, 1828. Pollen grains in water.
- Lord Kelvin, a stochastic approach for Boltzmann's equation, *Phil. Magazine*, 1901.
- ...

Monte Carlo methods are broadly applicable to different problems that are

- impossible to solve analytically
- difficult to handle for other numerical techniques due to the large computational complexity (dimension)

7.1.1 Buffon's Needle Experiment I

Buffon's (1707-1788) needle experiment to determine π:

- throw randomly needles, length l, on a grid of lines distance t apart
- what is the probability that the needle will lie across a line?

Define i.i.d random variables, x and α, are independent, use joint probability density function i.e.

$$\begin{cases} \frac{4t}{\pi} : & 0 \leq x \leq \frac{t}{2}, 0 \leq \alpha \leq \frac{\pi}{2} \\ 0 & \textit{else}. \end{cases}$$

The needle crosses a line if

$$x \leq \frac{l}{2} \leq \sin \alpha.$$

Integrating the joint probability density function gives the probability that the needle will cross a line.

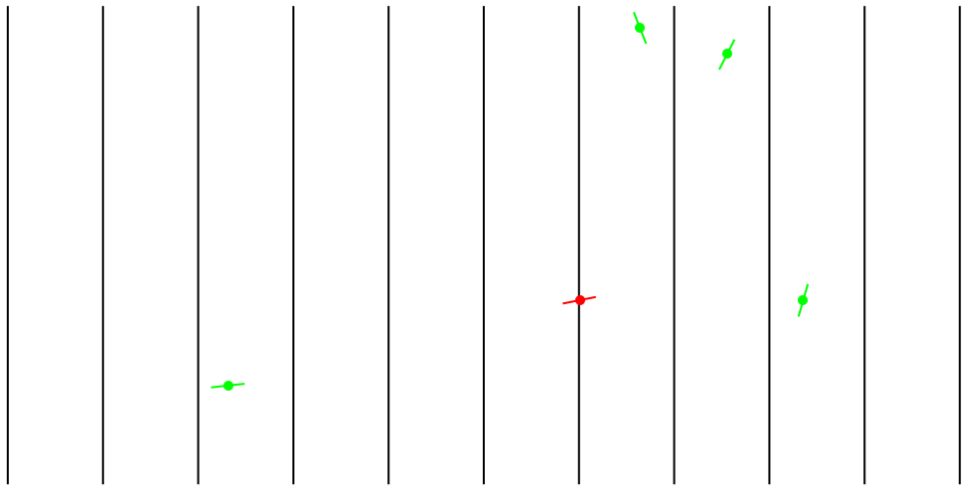
7.1.1 Buffon's Needle Experiment II

Buffon's (1707-1788) needle experiment to determine π:

Lazzarini 1901: Buffon's experiment with 34080 throws:

$$\pi \approx \frac{355}{113} = 3.14159292$$

A super nice result coming out of passion or curiosity driven research!

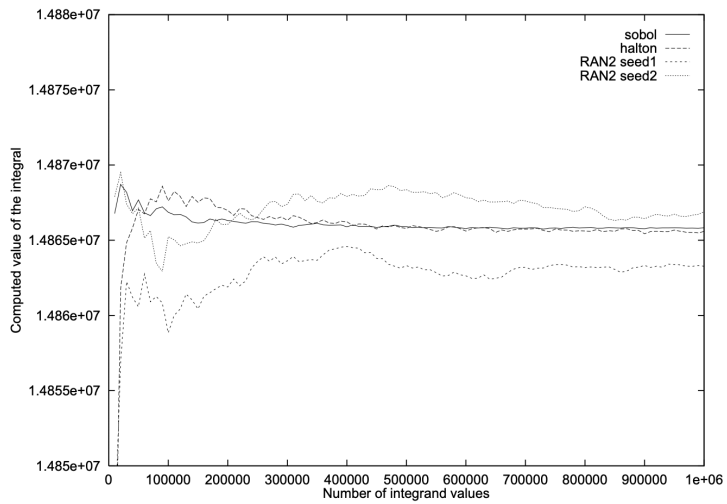


7.1.2 Computational Finances I

In 1995 Traub and Paskov computed values of 'mortgage-backed obligations' - in the US mortgages last for 30 years, and may be repaid each month, making $30 \times 12 = 360$ repayment possibilities.
The computed quantity is a 360-dimensional expected value of the form

$$\langle f \rangle = \int_0^1 \cdots \int_0^1 f(x^1, \ldots, x^{360}) dx^1 \cdots \ dx^{360}$$

Some of the conclusions from their paper (http://www.cs.columbia.edu/~traub/cucs-030-96.pdf):



7.1.3 The "DNA" of MC Methods I

- it may be difficult to numerically explore the complete phase or parameter space of a certain model
- we could apply appropriate random sampling techniques and only explore relevant parts of phase space
- if the number of samples is large enough, the computed estimate converges towards the correct value

Main Steps of a Monte Carlo Method

1. randomly choose a new configuration in phase space
2. accept or reject the new configuration
3. compute the physical quantity and add it to the averaging procedure
4. goto 1

7.2 The Error in Simple Monte Carlo I

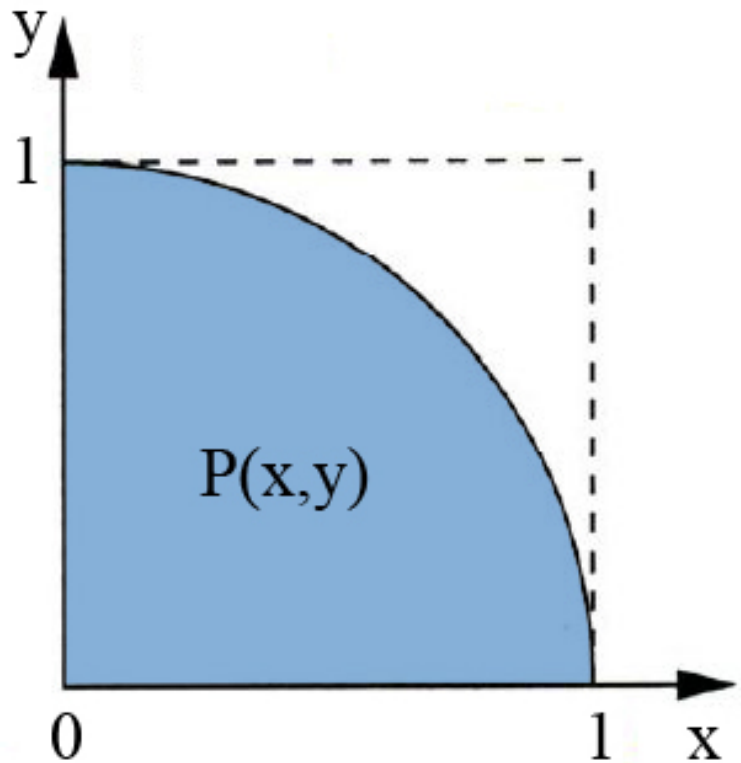
The Monte Carlo methods can be characterized by their reliance on repeated random sampling and averaging to obtain results. One of their major advantages is the following scaling of the error Δ with the number of samples N:

$$\Delta \propto \frac{1}{\sqrt{N}}$$

- physical and mathematical systems
- popular when an exact solution to a given problem cannot be found with a deterministic algorithm
- higher dimensional integration

7.3.1 Computation of π I

We consider the unit area ($x \in [0,1]$ and $y \in [0,1]$) and compare the area within the quarter circle, $P(x,y)$, to the area of the unit square ($P = \frac{\pi}{4}$). This relation is mathematically exact because $\pi = 4 \int_0^1 \sqrt{1-x^2}\, dx$.



7.3.1 Computation of π II

- consider N random points x_i, y_i in the unit square
- if $x_i^2 + y_i^2 \leq 1 \Rightarrow N_c = N_c + 1$

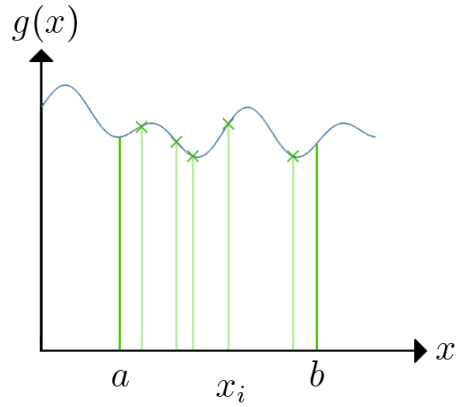
Then, the number of points N_c lying within the quarter circle is compared to the total number N of points and the fraction will give us an approximate value of π

$$\pi(N) = 4\frac{N_c(N)}{N}.$$

You will observe

$$\Delta = \pi(N) - \pi \propto \frac{1}{\sqrt{N}}.$$

7.4 Computation of Integrals I



- integral of a function $g(x)$ in an interval given by $[a, b]$
- approximate the integral by choosing N points x_i on the x-axis with their corresponding values $g(x_i)$
- summing and averaging over these sampled results and multiplying the resulting expression with the length of the interval

$$\int_a^b g(x)dx \approx (b-a)\left[\frac{1}{N}\sum_{i=1}^{N} g(x_i)\right]$$

7.4 Computation of Integrals II

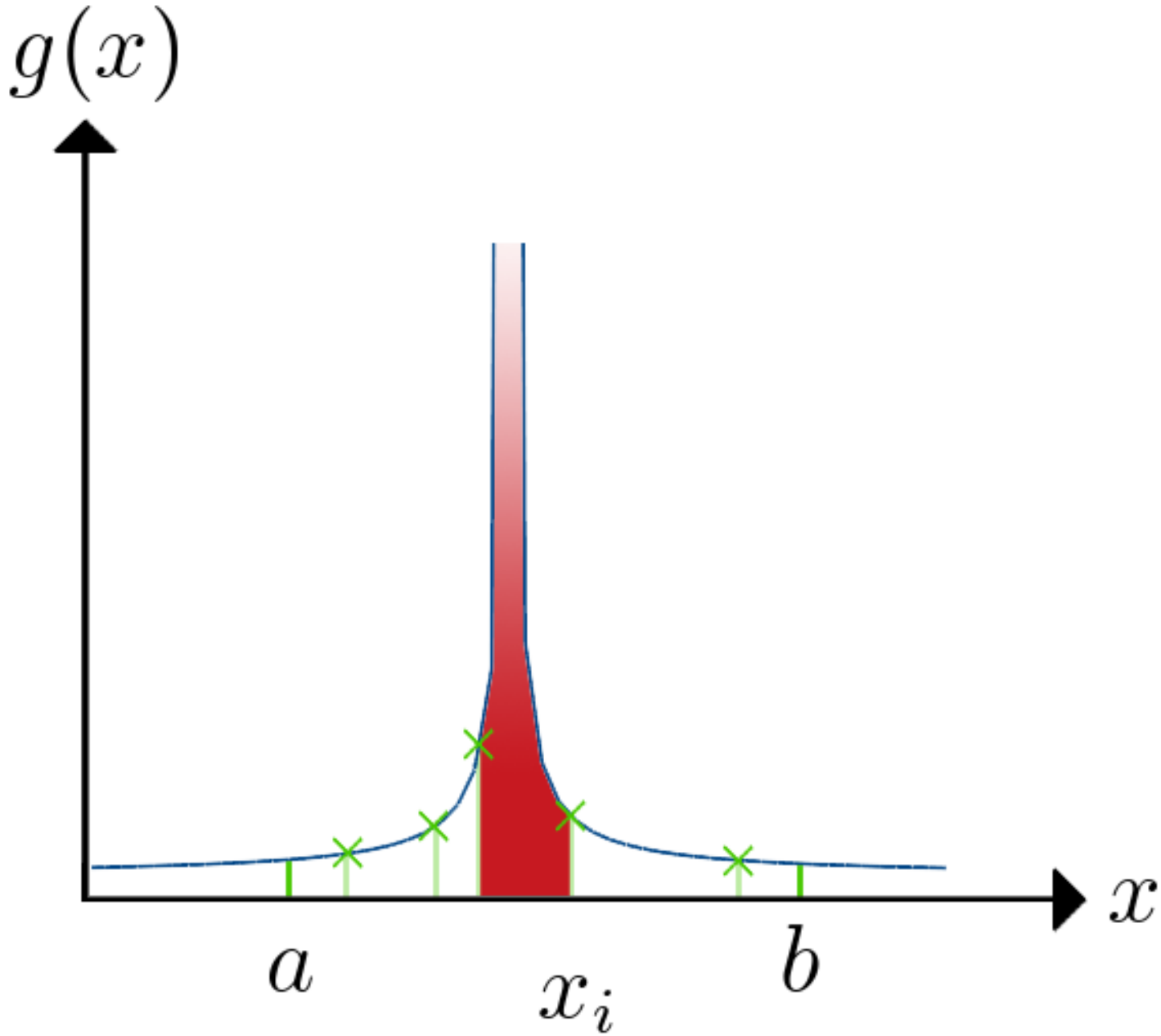
- what is the nature of these randomly chosen points x_i on the x-axis?
- if we choose them completely at random, the process is called "**simple sampling**" which works very well if $g(x)$ is smooth.

But if

- we cannot make the assumption that $g(x)$ is smooth?

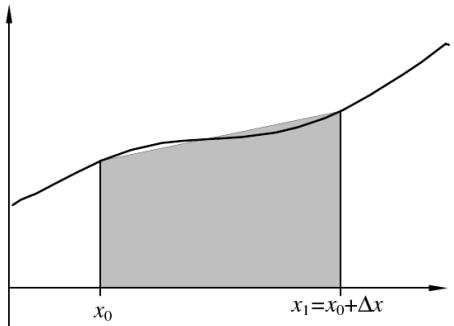
Let us consider a less "cooperative" function, for instance one featuring a singularity at a certain value. Due to the limited number of points that we would usually choose around the singularity (in simple sampling), we would not be able to really appreciate and understand its behavior. Furthermore, the integral would be a very rough approximation. We thus need more precision which goes beyond the possibilities of simple sampling and we need to refer to **importance sampling**.

7.4 Computation of Integrals III



7.4.1 Integration Errors I

Error in Conventional Methods - Trapezoidal Rule



Consider the Taylor Series expansion integrated from x_0 to $x_0 + \Delta x$:

$$\begin{aligned}
\int_{x_0}^{x_0+\Delta x} f(x)dx &= f(x_0)\Delta x + \frac{1}{2}f'(x_0)\Delta x^2 + \frac{1}{6}f''(x_0)\Delta x^3 + \ldots \\
&= \left[\frac{1}{2}f(x_0) + \frac{1}{2}(f(x_0) + f'(x_0)\Delta x + \frac{1}{2}f''(x_0)\Delta x^2 + \ldots) + \right. \\
&= \frac{1}{2}(f(x_0) + f(x_0+\Delta x))\Delta x + \mathcal{O}(\Delta x^3) \\
\rightarrow &\quad \text{local error } \propto (\Delta x)^3
\end{aligned}$$

7.4.1 Integration Errors II

Error in Conventional Methods - Trapezoidal Rule

We now subdivide our interval $[x_0, x_1]$ into N times $\Delta x = \frac{x_1 - x_0}{N}$.

$$\begin{aligned}
\int_{x_0}^{x_1} f(x)dx &\approx \frac{\Delta x}{2} \sum_{j=0}^{N-1} f(x_0 + j\Delta x) + f(x_0 + (j+1)\Delta x) \\
&= \frac{\Delta x}{2}[f(x_0) + 2f(x_0 + \Delta x) + 2f(x_0 + 2\Delta x) + \ldots \\
&\quad +2f(x_0 + (N-1)\Delta x) + f(x_1)] \\
&\rightarrow \text{cumulative error is } N \text{ times this or } \mathcal{O}((\Delta x)^2) \propto \mathcal{O}(N^{-2})
\end{aligned}$$

The generalization to d-dimensions is described in the script, the result is:

Dimension dependent error

The error in conventional methods $\propto (\Delta x)^2 \propto N^{-\frac{2}{d}}$

7.4.1 Integration Errors I

Monte Carlo Error

Consider a simplified case of a one dimensional function of one variable, $g : [a, b] \to \mathbb{R}$. If we pick N equidistant points in the interval $[a, b]$ we have a distance of $h = \frac{b-a}{N}$ between each of these points.
The estimate for the integral is

$$I = \int_a^b g(x)dx \approx \frac{b-a}{N} \sum_{i=1}^{N} g(x_i) = (b-a)\langle g \rangle \equiv Q$$

where $\langle g \rangle$ stands for the sample mean of the integrand.
The unbiased variance reads

$$\mathrm{Var}(g) \equiv \sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (g(x_i) - \langle g \rangle)^2$$

7.4.1 Integration Errors II

Monte Carlo Error

Using the Central Limit Theorem (CLT), the variance of the estimate of the integral can be computed as

$$\mathrm{Var}(Q) = (b-a)^2 \frac{\mathrm{Var}(g)}{N} = (b-a)^2 \frac{\sigma^2}{N}$$

which for large N decreases like $\frac{1}{N}$. Thus, the error estimate is

$$\delta Q \approx \sqrt{\mathrm{Var}(Q)} = (b-a)\frac{\sigma}{\sqrt{N}}$$

We can now generalize this to multidimensional integrals:

$$I = \int_{a_1}^{b_1} dx_1 \int_{a_2}^{b_2} dx_2 \ldots \int_{a_n}^{b_n} dx_n f(x_1, \ldots, x_n)$$

7.4.1 Integration Errors III

Monte Carlo Error

The previous interval $[a, b]$ becomes a hypercube V as integration volume, with

$$V = \{x : a_1 \leq x_1 \leq b_1, ...a_n \leq x_n \leq b_n\}$$

Instead of the interval $[a, b]$ we now use the hypercube V:

$$\mathrm{Var}(Q) = V^2 \frac{\mathrm{Var}(g)}{N} = V^2 \frac{\sigma^2}{N}$$

and the error estimate becomes

$$\delta Q \approx \sqrt{\mathrm{Var}(Q)} = V \frac{\sigma}{\sqrt{N}}$$

Dimension independent error

The error in simple Monte Carlo methods scales as $\frac{1}{\sqrt{N}}$ independent of the dimension d.

7.4.1 Integration Errors I

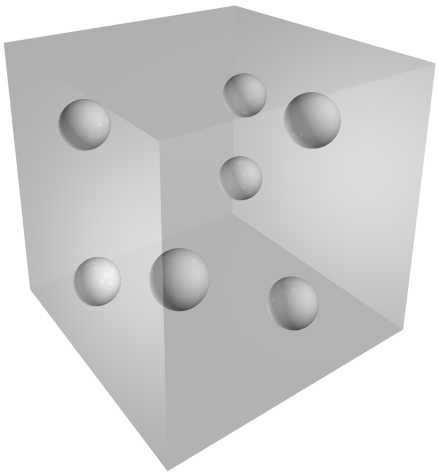
Comparison of the Errors

We have seen that in conventional methods, the error of computing integrals goes with $N^{-\frac{2}{d}}$ and thus depends on the dimension, while the error in Monte Carlo methods is independent of the dimension. There is a crucial point at which Monte Carlo methods become more efficient,

$$N^{-\frac{2}{d}} \overset{crit}{=} \frac{1}{\sqrt{N}} \qquad \Rightarrow d_{crit} = 4$$

We can thus conclude that for $d > 4$, Monte Carlo becomes more efficient and is therefore used in areas where such higher dimensional integrals with $d > 4$ are commonplace.

7.5 Higher Dimensional Integrals I



Let us now look at an example of higher dimensional integration: Consider N hard spheres of radius R in a 3D box of volume V. Our points are characterized by their position vector $\vec{x_i} = (x_i, y_i, z_i), 1 \leq i \leq N$. We define the distance between two such points as

$$r_{ij} := \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

where we have to take into consideration that the spheres are hard, i.e. cannot overlap. Thus the minimal distance between two neighboring spheres is the distance when the spheres are in contact, which is equal to $2R$. This translates to the following condition on r_{ij} :

$$r_{ij} \overset{!}{>} 2R$$

7.5 Higher Dimensional Integrals II

So far this does not seem to have too much to do with integration, we are simply placing some spheres in a volume V and measuring some distance r_{ij}. Let us say we are interested in the average distance between the centers of spheres in the box then we need to consider the following integral to reach an analytical result:

$$\langle r_{ij} \rangle = \frac{1}{Z} \int \frac{2}{N(N-1)} \sum_{i<j} r_{ij} d^3 r_1 ... d^3 r_N \quad , \text{ where } \quad Z = \int d^3 r_1 ... d^3 r_N$$

- the first factor, $\frac{1}{Z}$, is a normalization factor
- the second factor is of combinatory origin, in fact it stems from random drawing from a finite population without replacement.
- as they are indistinguishable, there is an additional factor of $\frac{1}{2}$

Note that we would need to correct with a factor of $\frac{1}{2}$ if the sum were over $i \neq j$ (since, in that case, we would have counted each volume twice) which was avoided by simply summing over $i < j$.

7.5 Higher Dimensional Integrals III

The Monte Carlo approach to this formula is relatively simple:

- Choose a particle position (i.e. the center of the new sphere)
- Make sure that the new sphere does not overlap with any pre-existing spheres (see condition on r_{ij}). If it does overlap, reject the position and try again
- Once all the spheres have been placed, calculate the distances r_{ij}.

We then use these distances r_{ij} to compute the average.

7.6 Utilizing Nonuniform Random Numbers I

Assumption: the random numbers are generated in a d-dimensional box.

In case the volume $V \subset \mathbb{R}^d$ to be integrated fills only a small fraction of that box, the fraction of misses can be enormous

- find some other distribution which better encloses the volume we are integrating over, and generate random numbers only in this
- if the enclosing function is such that we know how to generate random numbers in this distribution analytically, the savings in time can be enormous

7.6 Utilizing Nonuniform Random Numbers II

This is actually exactly what is done in the combined analytical-rejection method to generate random numbers.

1. generate a uniformly distributed number $u = P_u(0, 1)$
2. generate a number distributed as $g(x) : x = G^{-1}(u)$
3. generate a uniformly distributed number $y = P_u(0, g(x))$
4. If $y > f(x)$ this is a **miss**: goto 1
5. else **hit** return x

$\Rightarrow$ improving the MC accuracy $\Rightarrow$ reducing the variance σ^2

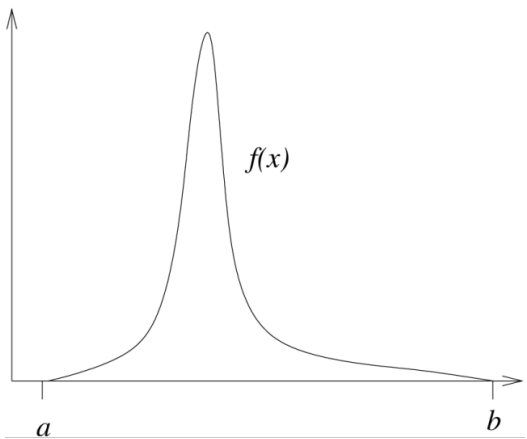
Remember that σ^2 for any non-constant data distribution is a quantity which goes towards some finite, non-zero value when the number of sample $N \to \infty$, whereas the error of course goes to 0 with increasing N.

7.6.1 Importance Sampling I

Idea

Since the MC error is $\Delta \propto \frac{1}{\sqrt{N}}$, it is clear that if we can reduce the variance σ^2 the error will also go down for the same N.

It is obvious that most of the integral comes from the region of the peak. But if we generate points evenly in the interval $[a, b]$, most points won't be in the peak area, and their contribution to the total will be relatively small.



7.6.1 Importance Sampling II

In fact, some simply thought indicates that the least effort will be spent in case the distribution is fairly flat. In that case the variance σ^2 will become smaller.
The idea behind importance sampling:

- is to transform $f(x)$ into another, flatter function
- inverse transformation must exist

Assume we have a function $g(x)$ normalized over the integration interval $D = [a, b]$ which gives the property that

$$\frac{f(x)}{g(x)}, \quad g(x) > 0, \forall x \in D.$$

is fairly flat.
We now want to calculate

$$I = \int_a^b f(x)dx = \int_a^b \frac{f(x)}{g(x)} g(x)dx = \int_a^b \frac{f(x)}{g(x)} dG(x)$$

7.6.1 Importance Sampling III

where

$$G(x) = \int_a^x g(x)dx \tag{1}$$

is the integral of $g(x)$.
We make a variable change $r = G(x)$ and we get

$$I = \int_{G(a)}^{G(b)} \frac{f(G^{-1}(r))}{g(G^{-1}(r))} dr \tag{2}$$

which gives the same integral, but more efficiently because the integrand is flatter than the original.
Evaluating this with MC is done in the same way as for any other function,

$$I \sim \frac{1}{N} \sum_{i=1}^{N} \frac{f(G^{-1}(r_i))}{g(G^{-1}(r_i))} \tag{3}$$

where the r_i are uniform random numbers.

7.6.1 Importance Sampling IV

So we have to be able to determine G^{-1}. However there is an alternative: it is actually enough that we can generate random points distributed as $g(x)$ by any means (not necessarily analytically). In this case the MC sum is

$$I \sim \frac{1}{N} \sum_{i=1}^{N} \frac{f(x_i^G)}{g(x_i^G)} \tag{4}$$

where the x_i^G are **random numbers distributed** as $g(x)$.

7.6.2 Control Variates I

- similar idea: variance reduction
- use flatter function

But instead of division use subtraction:

$$I = \int_a^b f(x)dx = \int_a^b (f(x) - g(x))dx + \int_a^b g(x)dx \tag{5}$$

The idea is to find a $g(x)$ such that

Idea

- $\mathrm{Var}(f - g) < \mathrm{Var}(f)$
- $\int_a^b g(x)dx$ is known

The integral $\int_a^b (f(x) - g(x))dx$ is evaluated with normal MC sampling. The following advantages compared to importance sampling are obvious:

7.6.2 Control Variates II

- $g(x)$ can be zero somewhere in the interval $[a, b]$
- we do not need to know how to distribute random numbers according to $g(x)$

Now the term $\mathrm{Var}(f - g)$ is

$$\mathrm{Var}(f - g) = \mathrm{Var}(f) + \mathrm{Var}(g) - 2\mathrm{Cov}(f, g) \tag{6}$$

where the last term is the covariance between f and g. In order that this be useful, we have the constraint that

$$2\ \mathrm{Cov}(f, g) > \mathrm{Var}(g). \tag{7}$$

which translates into: f and g need to be positively correlated.

7.7 Quasi-Monte Carlo I

$$Q = \frac{1}{N} \sum_{i=1}^{N} f(x_n) \tag{8}$$

with $x_1, \ldots, x_n$ deterministic (and cleverly chosen). How to choose $x_1, \ldots, x_n$?

- lattice rules (many of them exist)
- low discrepancy points (see chapter on random numbers)

For $V < \infty$ the approximation properties of deterministic low discrepancy sequences scales as

$$\mathcal{O}\left(\frac{(\log N)^d}{N}\right).$$

- It has been noted that the convergence rate of quasi-Monte Carlo method in practice is usually much faster than its theoretical bound [1].

7.7 Quasi-Monte Carlo II

- this advantage is not universal

Some obvious drawbacks are immediately visible:

- for $\mathcal{O}\left(\frac{(\log N)^d}{N}\right) < \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \Rightarrow N > 2^d$
- for large d and practical N values, the discrepancy D^* of a point set from a low-discrepancy generator might be not smaller than for a random set.
- for many functions arising in practice, $V(f) = \infty$

In order to overcome some of these deficiencies, we can use a randomized quasi-Monte Carlo method however, this is out of the scope of this introduction.

[1] Soren Asmussen and Peter W. Glynn, Stochastic Simulation: Algorithms and Analysis, Springer, 2007—

Interlude Markov Chains I

Idea

A Markov Chain is a stochastic (always changing) model that is used to predict/estimate/guess the outcome of an event given only the previous state and its action.
If a sequence of events exhibits the Markov Property of the reliance on the previous state, then the sequence is called "Markovian" in nature.

Probelm setup: Suppose that we are interested in generating samples from a target probability distribution π on $\mathbb{R}^n$ that π is such that (complex) we can not use direct methods for simulation. Using Markov chain Monte Carlo methods it is, however, often feasible to generate an ergodic Markov chain $X_1, X_2, \ldots$ which has π as equilibrium distribution, i.e. after a suitable burn-in period $m, X_{m+1}, X_{m+2}, \ldots$ provides a (correlated) sample from π which can be used e.g. for Monte Carlo computations.

Interlude Markov Chains II

A Markov chain X is a sequence $X_1, X_2, \ldots$ of stochastic variables, which for all $n > 0$ and all events $A_1, A_2, \ldots, A_n$ satisfies the following conditional independence property:

$$P(X_n \in A_n | X_{n-1} \in A_{n-1}, \ldots, X_0 \in A_0) = P(X_n \in A_n | X_{n-1} \in A_{n-1})$$

As a first specific example we consider a Markov chain on the discrete state space $E = \{\text{Ground State}, \text{Exited State}\} = \{G, E\}$.
A Markov chain X on $E = \{G, R\}$ is determined by the initial distribution given by

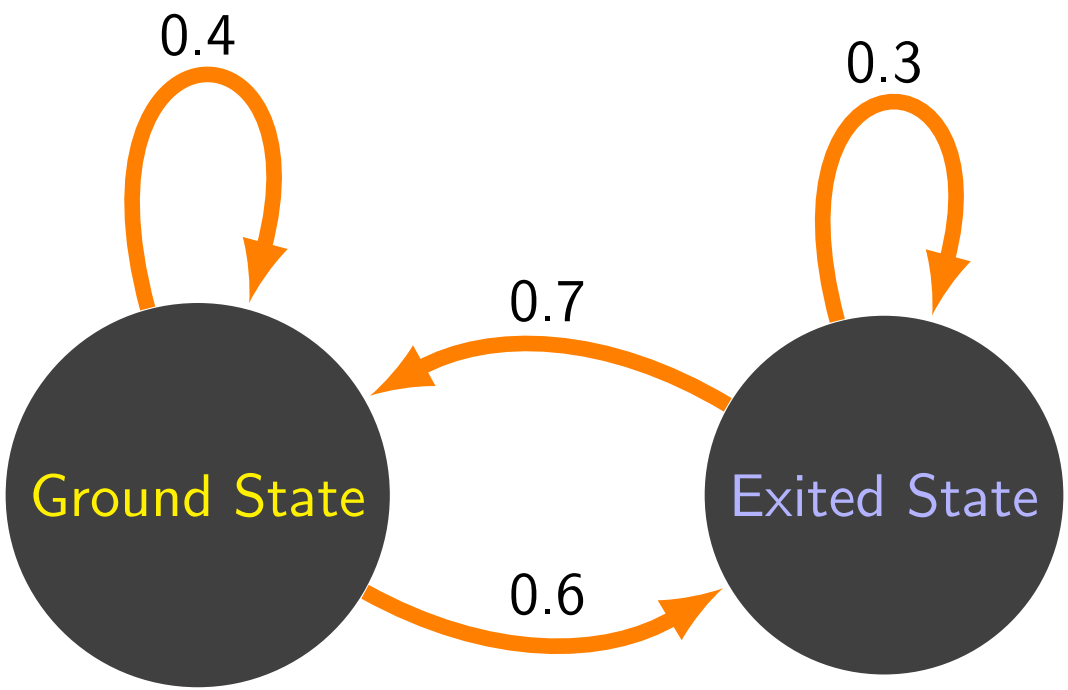
$$p_0 = P(X_0 = G) \text{ and } p_1 = P(X_0 = E)$$

and the one-step transition probabilities given by

$$p_{00} = P(X_{n+1} = G | X_0 = G), \quad p_{10} = P(X_{n+1} = G | X_0 = E)$$

Interlude Markov Chains III

$$p_{01} = 1 - p_{00}, \quad p_{11} = 1 - p_{10}$$



Interlude Markov Chains IV

The one-step transition probabilities can be written as a matrix

$$P = \begin{pmatrix} p_{00} & p_{10} \\ p_{01} & p_{11} \end{pmatrix} = \begin{pmatrix} 0.4 & 0.6 \\ 0.3 & 0.7 \end{pmatrix}$$

The next example is a Markov chain with the continuous state space $E = \mathbb{R}$.

7.8 Canonical Monte Carlo I

What is an Ensemble?

- an ensemble is a set of a large number of identical systems
- one usually considers the phase space (which for N particles is $6N$ dimensional)
- elected points are then regarded as a collection of representative points in phase space.

An important and ever-recurring topic is the probability measure:

- the statistical properties depend on the probability measure
- assume we have two regions R_1 and R_2 with R_1 having a larger measure than R_2.
- if we pick a system at random from our ensemble, it is more probable that it is in a microstate pertaining to R_1 than R_2.

Two more terms:

7.8 Canonical Monte Carlo II

What is an Ensemble?

- the normalizing factor of the measure is called the partition function of the ensemble
- an ensemble is said to be stationary if the associated measure is time-independent.

The most important ensembles are the following:

- Microcanonical ensemble: A closed system with constant number of particles N, constant volume V and constant inner energy E
- Canonical ensemble: A closed system in a heat reservoir with constant N, constant temperature T and constant V
- Grand canonical ensemble: An open system where the chemical potential μ is constant along with V and T.

7.8 Canonical Monte Carlo III

What is an Ensemble?

The ensemble average of an observable (i.e. a real-valued function f) defined on phase space Λ with the probability measure $d\mu$ (restricting to μ-integrable variables) is defined by

$$\langle f \rangle = \int_{\Lambda} f d\mu$$

The time average is defined in a different way. We start out with a representative starting point $x(0)$ in phase space. Then, the time average of f is given by

$$\bar{f}_t = \lim_{T \to \infty} \frac{1}{T} \int_0^T f(x(t)) dt$$

These two averages are connected by the Ergodic hypothesis. This hypothesis states that for long periods of time, the time one particle spends in some region Λ of phase space of microstates with the same

7.8 Canonical Monte Carlo IV

What is an Ensemble?

energy is directly proportional to the volume $V(\Lambda)$ of that region. One can thus conclude that all the accessible microstates are equiprobable (over a long period of time). In statistical analysis, one often assumes that the time average $\bar{Q}_t$ of some quantity Q and the ensemble average $\langle Q \rangle$ are the same.

7.8 Back to Canonical Monte Carlo I

Let the energy of configuration X be given by $E(X)$, then the probability (at thermal equilibrium) for a system to be in X is given by the Boltzmann distribution:

$$p_{eq}(X) = \frac{1}{Z_T} e^{-\frac{E(X)}{k_B T}} \text{ with } \sum_X p_{eq}(X) = 1$$

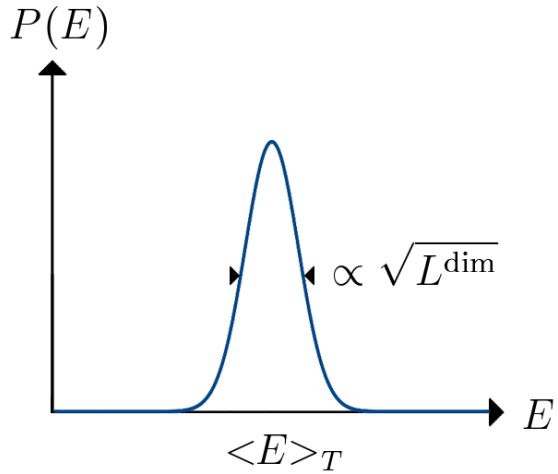
and Z_T is the partition function (the normalizing factor of the measure) :

$$Z_T = \sum_X e^{-\frac{E(X)}{k_B T}}$$

Let us now consider an ensemble average which for a discrete X reads

$$\langle Q \rangle = \sum_X Q(X) p_{eq}(X)$$

7.8 Back to Canonical Monte Carlo II



- calculate the ensemble average for a given property $\rightarrow$ energy
- the distribution of energy around the time average $\bar{E}_t$ gets sharper with increasing system size (the peak width increases with the system size as $\sqrt{L^{dim}}$ while the system increases with L^{dim} so the relative width decreases as $\frac{1}{\sqrt{L^{dim}}}$)so the relative width decreases as $\frac{1}{\sqrt{L^{dim}}}$)

7.8 Back to Canonical Monte Carlo III

Consequently, it is inefficient to pick equally distributed configurations over the energy. We are thus still (at least) one central ingredient short for the computation of the ensemble average.

- a common way to efficiently choose appropriate samples out of a large pool of possible configurations is to explore the phase space using a Markov chain.
- importance sampling

Setup

- introduce the virtual time τ representing the steps of a stochastic process
- start in a given configuration X and propose a new configuration Y with a transition probability $T(X \to Y)$

7.8 Back to Canonical Monte Carlo IV

- normalization: over all possible new configurations to be unity $\sum_Y T(X \to Y) = 1$
- reversibility: $T(Y \to X)$
- from thermodynamics we need the ergodic hypothesis, which states that thermodynamical systems are generally such as volume elements corresponding to a given energy are equiprobable.

In the light of importance sampling: not every new configuration Y will be accepted.

7.8 Back to Canonical Monte Carlo I

Markov Chain Probability (using the Ising model)

- we may propose a new configuration that will increase the energy of the system, but depending on the temperature it is not necessarily likely that the new configuration will also be accepted
- at low temperatures, it is unlikely that a spin flip will occur, so we will handle this with an acceptance probability A, $A(X \to Y)$

In practice, we might be interested in the overall probability of a configuration actually making it through these two steps; this probability is the product of the transition probability $T(X \to Y)$ and the acceptance probability $A(X \to Y)$.

$$W(X \to Y) = T(X \to Y) \cdot A(X \to Y)$$

If we are interested in the evolution of the probability $p(X, \tau)$ to find X in time, we can derive this with a logical Gedankenexperiment: there are two processes changing this probability:

7.8 Back to Canonical Monte Carlo II

Markov Chain Probability (using the Ising model)

- A configuration X is reached by coming from $Y \Rightarrow \nearrow p(X, \tau))$
- A configuration X is left, goto $Y \Rightarrow \searrow p(X, \tau)$

The first of these two is proportional to the probability for a system to be in Y, $p(Y)$, while the second one needs to be proportional to the probability for a system to be in X, $p(X)$: **Master equation** :

$$\frac{p(X,\tau)}{d\tau} = \sum_{Y \neq X} p(Y) W(Y \to X) - \sum_{Y \neq X} p(X) W(X \to Y)$$

The properties of $W(X \to Y)$ are:

1. Ergodicity: $\forall X, Y : W(X \to Y) > 0$
2. Normalization: $\sum_Y W(X \to Y) = 1$
3. Homogeneity: $\sum_Y p(Y) W(Y \to X) = p(X)$

7.8 Back to Canonical Monte Carlo III

Markov Chain Probability (using the Ising model)

Theorem

A Markov chain that satisfies conditions $1-3$ *converges towards a unique stationary distribution* p_{st}.

To see this, we set the stationary distribution p_{st} equal to the equilibrium distribution of the considered physical system p_{eq}

$$\frac{dp(X,\tau)}{d\tau} = 0 \Leftrightarrow p_{st} \stackrel{!}{=} p_{eq}.$$

Detailed Balance

This leads to the concept of *Detailed Balance* and tells us that the steady state of the Markov process in the thermal equilibrium. We have achived that by using the Boltzman distribution for $p(X)$.

Details can be found in the script.

7.8.1 M(RT)^2 Algorithm I

- M(RT)^2 is an abbreviation of the last names of the authors of the original paper
- RT is squared because except Metropolis, the other four authors (Rosenbluth, Teller) of the paper formed two married couples
- The real contributions of some of the authors (in particular of Metropolis and of A.H. Teller) is subject of controversy
- It has been even stated that the original algorithm was invented by Enrico Fermi

7.8.1 M(RT)^2 Algorithm II

In the Metropolis algorithm, the acceptance probability A is defined as

$$A(X \to Y) = \min\left(1, \frac{p_{eq}(Y)}{p_{eq}(X)}\right)$$

We can now insert the Boltzmann distribution

$$p_{eq}(X) = \frac{1}{Z_T} e^{-\frac{E(X)}{k_B T}}$$

to find

$$A(X \to Y) = \min\left(1, e^{-\frac{E(Y)-E(X)}{k_B T}}\right) = \min\left(1, e^{-\frac{\Delta E}{k_B T}}\right)$$

Suppose we go to a configuration of lower energy, ΔE will be negative and we would end up with an exponential expression like $\exp(\frac{|\Delta E|}{k_B T})$, at which point the minimum kicks in and sets the expression to one. In other words,

7.8.1 M(RT)^2 Algorithm III

the acceptance will be equal to one ("always accept") for transitions to configurations of lower energy.
If we go to a configuration of higher energy, ΔE will be positive and we'll have $\exp(-\frac{|\Delta E|}{k_B T}) < 1$ so the acceptance will increase with the temperature.

Remark

Note that a thermal equilibrium is enforced by detailed balance and we impose that the steady state must be a Boltzmann distribution.

7.8.1 M(RT)^2 Algorithm IV

The M(RT)^2 Algorithm

1. Randomly choose a configuration X_i
2. Compute $\Delta E = E(Y) - E(X)$
3. Spinflip if $\Delta E < 0$, otherwise accept with prob. $\exp\left(-\frac{\Delta E}{k_B T}\right)$

7.9 The Ising Model I

- the Ising model is a model that originally aimed at explaining ferromagnetism
- today used in many other areas such as opinion models, binary mixtures, lattice gas ...
- a highly simplified approach to e.g. the difficulties of magnetism (e.g. commutation relations of spins)

Setup:

- we consider a discrete collection of N binary variables called spins
- they take on the values ± 1 (representing the up and down spin configurations)
- the spins σ_i interact pairwise
- the energy has one value for aligned spins ($\sigma_i = \sigma_j$) and another for anti-aligned spins ($\sigma_i \neq \sigma_j$)

7.9 The Ising Model II

The Hamiltonian is given by

$$\mathcal{H} = E = -\sum_{ij} J_{ij}\sigma_i\sigma_j - \sum_i H_i\sigma_i \tag{9}$$

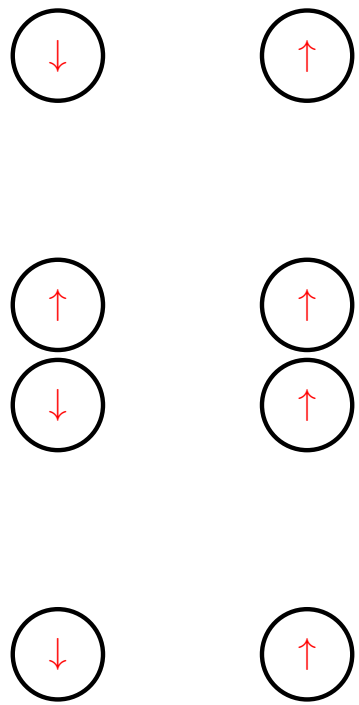
where H_i is a (usually homogeneous) external field and J_{ij} are the (translationally invariant) coupling constants. As the coupling constants are translationally invariant, we may drop the indices and simply write $J = J_{ij} \ \forall i, j$. The coupling constant J is half the difference in energy between the two possibilities (alignment and anti-alignment).

The simplest example is the antiferromagnetic one-dimensional Ising model, which has the energy function

$$E = \sum_i \sigma_i\sigma_{i+1} \tag{10}$$

7.9 The Ising Model III

which can be generalized to the two-dimensional case; we can also add an external field as in equation (10).
Consider for illustration an ensemble of 4 particles with spins i.e., $|S| = 2^4 = 16$ states $X_1, \ldots, X_{|S|}$. A particular state $X_j = (\downarrow, \uparrow, \uparrow, \uparrow)$ is depicted below.



7.9 The Ising Model IV

A Markov chain is generated by stepping from X_i to X_{i+1}, with probability $p_{i,i+1}$ or short $p_{i,j}$ with $j = i + 1$. The probability $p_{i,j}$ is called the transition probability. With $\boldsymbol{P}$ the transition matrix comprised of $p_{i,i+1}$

$$p_{i,j} = X_i^T \boldsymbol{P} X_j, \quad \dim \boldsymbol{P} = |S| \times |S|. \tag{11}$$

The following normalizing condition must be fulfilled:

$$p_i = \sum_{j=1}^{|S|} p_{i,j} = 1 \tag{12}$$

were p_i means the probability to any state i.
A Markov process creates a sequence of states $X^{(\tau)}$ that are labeled by the *virtual/algorithmic time* (τ).

7.9 The Ising Model V

The M(RT)^2 Algorithm & the Ising Model

1. Randomly choose a configuration X_i
2. Compute $\Delta E = E(Y) - E(X) = 2J\sigma_i\sigma_j$
3. Spinflip if $\Delta E < 0$, otherwise accept with prob. $\exp\left(-\frac{\Delta E}{k_B T}\right)$

7.9 The Ising Model I

Some Details in 1D