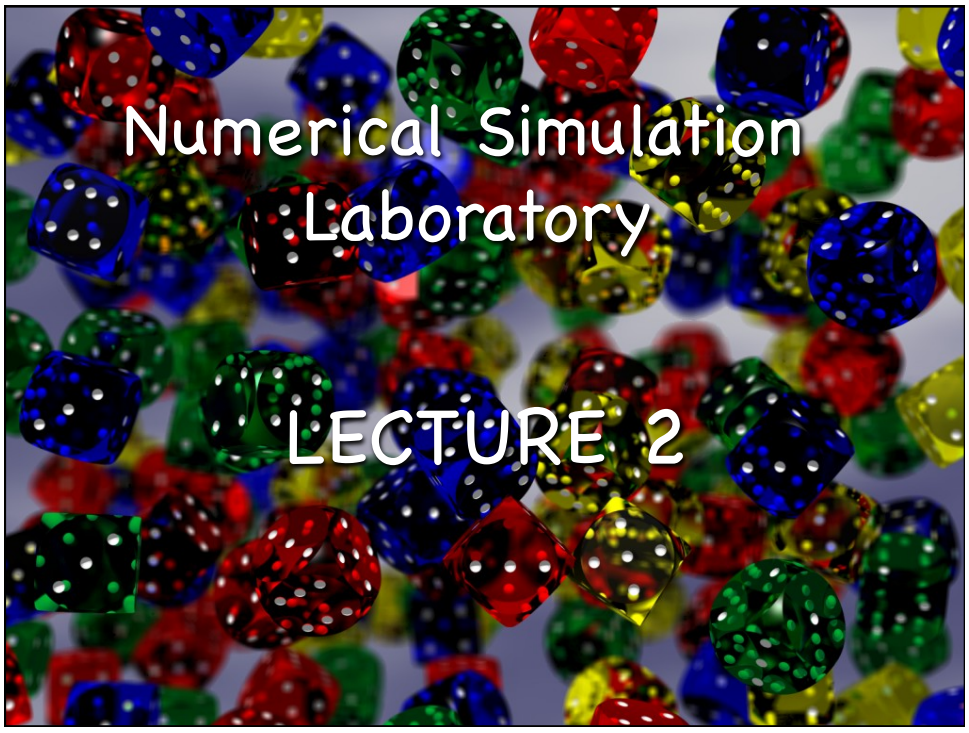


Lecture 2 - outline

- Sampling discrete distributions
- Monte Carlo integration
- Importance sampling
- Rejection techniques
- Stochastic processes
- Markov processes
- Random walk on a lattice

Sampling discrete distributions

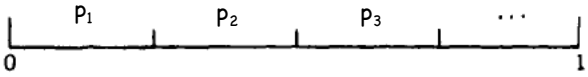
- The probability distribution function for a uniform random variable r is $p_r(r)=1$, $0 \le r < 1$; we have that

$$P[0 \le x_1 < r \le x_2 \le 1] = F_r(x_2) - F_r(x_1) = x_2 - x_1$$

  The chance that r lies in an interval $[x_1, x_2]$ of [0,1] is equal to the length of the interval.

- Suppose we have a class of events E_k with probabilities p_k and we wish to sample one at random.

- Since $\Sigma_k p_k = 1$, it is possible to take the interval [0,1] and exhaust it by dividing it into segments each of which has a length equal to some p_l. Then you can sample a uniform random number $r \in [0,1)$ ...



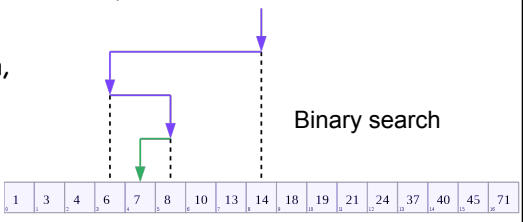
The interval into which a r falls determines the identity of the event.

- A uniform random variable is generated, and the smallest l is found for which the sum of the p_k is greater than the random number; that is,

$$\sum_{k=0}^{l-1} p_k < r \le \sum_{k=0}^{l} p_k$$

  (When l=0, the sum is defined to equal 0.) Whenever $0<r<p_1$ event 1 takes place; if $p_1<r<p_1+p_2$ event 2 takes place; and so on.

- In searching for an index l satisfying the previous relation, a binary search is strongly recommended when the total number of intervals is large and if $\Sigma_k p_k$ does not converge fast. If a serial search is to be used and the index can be arranged at our disposal, then the index with the largest probability should be put in the first place and so on, to reduce the average time of searching.



- Find many other cases (i.e., how to sample different one-dimensional probability distributions) in Bratley, Fox, Schrage *A Guide to Simulation* - Springer 1987.

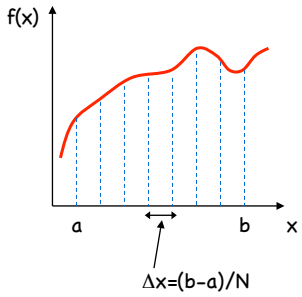
Monte Carlo integration

- Let us consider the following integral:

$$I = \int_a^b f(x)dx$$

and a partition of the interval [a,b] ; the value I can be obtained with

$$I = \int_a^b f(x)dx = \lim_{N\to\infty} \sum_{i=1}^{N} \Delta x \cdot f(x_i)$$



- Previous equation can be rewritten as

$$I = \lim_{N\to\infty} \sum_{i=1}^{N} \Delta x \cdot f(x_i) = (b-a) \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} f(x_i) = (b-a)\langle f \rangle_{[a,b]}$$

an estimation of the mean of f, ⟨f⟩, can be obtained by considering random variables x_i uniformly distributed in the interval [a,b]:

$$I = (b-a)\langle f \rangle_{[a,b]} \cong (b-a)\frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

Monte Carlo evaluation of integrals

- We now explore somewhat more systematically the ideas that underlie Monte Carlo quadrature. If an integral must be evaluated having the form

$$I = \int_\Omega g(x)p(x)dx$$

where $p(x) \geq 0 \; \forall x \in \Omega$; $\int_\Omega p(x)dx = 1$

then the following "game of chance" may be used to make numerical estimates. We draw a set of variables $x_1,\ldots,x_N$ from p(x) [i.e., we "sample" the probability distribution function p(x)] and form the arithmetic mean

$$G_N = \frac{1}{N}\sum_{i=1}^{N} g(x_i)$$

- The quantity G_N is an estimator for I and we have seen that $\langle G_N \rangle = I$ if the integral exists. Since G_N estimates I, we can write $G_N = I + \text{error}$. If the variance exists, the error appearing in the last statement is a random variable whose width is characterized for large N by

$$error = \varepsilon = \frac{\sigma_I}{\sqrt{N}}$$

where

$$\sigma_I^2 = \int_\Omega g^2(x)p(x)dx - I^2$$

- The error estimate may be inverted to show the number of samples needed to yield a desired error, ε:

$$N = \sigma_I^2 / \varepsilon^2$$

- The integral to be done need not exhibit explicitly a function p(x) satisfying the properties of a probability distribution. One can simply use p(x)=1/Ω and g(x)= Ω·integrand.
- The integral I could also be evaluated by standard quadrature. Let us assume that the domain of integration is an n-dimensional unit hypercube; then a numerical integration procedure can be written:

$$I \cong \sum_{i=1}^{N} w_i g(x_i) p(x_i)$$

where x_i is a lattice of points that fills the unit hypercube and w_i is a series of quadrature weights.

- The error associated with this quadrature is bounded by $\varepsilon \le ch^k$ where h measures the size of the interval separating the individual x_i .
- The constants ***c*** and ***k*** depend on the actual numerical integration method used, and ***k*** normally increases with more accurate rules. The bound on the error is not a statistical variable, but is an absolute number.

- If we assume that the time necessary for a computation will be proportional to the total number of points used, then

$$t_c \propto N = \gamma \left(1/h\right)^d$$

where γ is a constant of the order of 1 and d is the number of dimensions.

- The equation with the bound for the error can be rewritten

$$h \ge \left(\varepsilon / c\right)^{1/k}$$

- and the time t_c becomes

$$t_c \propto \gamma \left(c/\varepsilon\right)^{d/k} = t_0 \varepsilon^{-d/k}$$

- The greater the accuracy demanded in a calculation, the greater the computational time will be.
- In a Monte Carlo calculation, the total computation time is the product of the time for an individual sampling of x, t_I , times the total number of points: $t_c = t_I N$
- Using the previously estimated N, this may be rewritten as

$$t_c = t_I \sigma_I^2 / \varepsilon^2 = Q_I \varepsilon^{-2}$$

the exponent of ε is the same in any number of dimensions.

- For large d it is difficult to find a k in the equation for t_c such that $d/k < 2$, so asymptotically ($n \to \infty$) a Monte Carlo calculation is more advantageous than a numerical integration. The Monte Carlo calculation will take less total time for the same value of the error ε. This assumes that the two error estimates can be directly compared.
- In spite of the apparently slow convergence ($\approx N^{-1/2}$) of the error of Monte Carlo quadrature, it is in fact more efficient computationally than finite difference methods in dimensions higher than six to ten.
- Two different Monte Carlo evaluations of an integral can have differing variances. The quantity

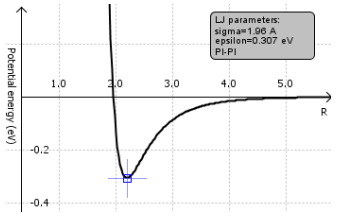
$$Q_I = t_I \sigma_I^2$$

  is a measure of the quality (efficiency) of a Monte Carlo calculation.
- The decision on which Monte Carlo algorithm to use in a large computation can be based on the values of Q_I extracted from some trial calculations.
- A common phenomenon is for t_I to increase as σ_I decreases through a more elaborate Monte Carlo algorithm. The question is then whether the decrease in σ_I will more than compensate for the increase in time. It will if Q_I decreases.

Even more... an example

- System: N (for example 100) interacting particles in a box

  pair-additive interacting potential v(|r|) →



- Hamiltonian: $H = \sum_{i=1}^{N} \frac{\vec{p}_i^2}{2m} + \sum_{i<j=1}^{N} v\left(\left|\vec{r}_i - \vec{r}_j\right|\right)$
- Statistical mechanics:

  Canonical (NVT) ensemble, Gibbs statistical weight (configurational probability density):

$$p(\vec{r}_1,\cdots,\vec{r}_N) = \frac{e^{-\frac{1}{k_B T} V(\vec{r}_1,\cdots,\vec{r}_N)}}{Z} = \frac{e^{-\frac{1}{k_B T}\sum_{i<j=1}^{N} v\left(\left|\vec{r}_i - \vec{r}_j\right|\right)}}{Z}$$

- Typical average: mean of potential energy:

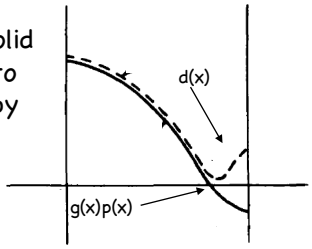
$$\langle V \rangle = \iiint d\vec{r}_1 \cdots d\vec{r}_N p(\vec{r}_1,\cdots,\vec{r}_N) V(\vec{r}_1,\cdots,\vec{r}_N)$$

- Classical quadrature: estimation of the integrand on n=10 points for each dimension (3N) ⇒ total number of points: $n^{3N}=10^{300}$ ... NEVER!!!
- MONTE CARLO: 3N=300 dimension is not a problem (daily I integrate on about 10^5 dimensions!!!), but we have to learn how to sample $p(r_1,...,r_N)$
- The Metropolis algorithm will be the answer!

Importance sampling

- Suppose we have an n-dimensional integral that we wish to evaluate. $$I = \int g(\vec{x})p(\vec{x})d\vec{x}$$
- The function p(x) is not necessarily the best probability distribution to use in the Monte Carlo calculation even though it appears in the integrand. A different probability distribution , can be introduced into the integral as follows:
$$I = \int \left[\frac{g(\vec{x})p(\vec{x})}{d(\vec{x})}\right] d(\vec{x})d\vec{x}$$
where $d(\vec{x}) \geq 0 \; \forall \vec{x}$, $\int d(\vec{x})d\vec{x} = 1$
and g(x)p(x)/d(x)<∞ except on a set of points with zero measure.
- The variance of I when d(x) is used becomes
$$\mathrm{var}[I]_d = E\left[\left(\frac{gp}{d}\right)^2\right]_d - \left(E\left[\frac{gp}{d}\right]_d\right)^2 = \int \left[\frac{g(\vec{x})p(\vec{x})}{d(\vec{x})}\right]^2 d(\vec{x})d\vec{x} - I^2$$
I being fixed, we want the d(x) that will minimize the $E[(gp/d)^2]$.
- Of course, the integral is minimized by choosing d(x) as large as we like, but we have the additional constraint expressed by the normalization of the probability distribution d(x).

- The function d(x) that satisfies the criteria given above may be deduced by using a Lagrange multiplier λ. In this method we wish to find d(x) such that
$$L\{d\} = \left[\int \left[\frac{g(\vec{x})p(\vec{x})}{d(\vec{x})}\right]^2 d(\vec{x})\, d\vec{x} + \lambda\left(\int d(\vec{x})\, d\vec{x} - 1\right)\right]$$
is minimized.
- We consider small variations of d(x) on the quantity in brackets and set the variation of the quantity in brackets equal to zero
$$\frac{\delta}{\delta d}[\cdots] = 0$$
- Performing the functional differentiation yields: $$-\left[\frac{g(\vec{x})p(\vec{x})}{d(\vec{x})}\right]^2 + \lambda = 0$$
or $$d(\vec{x}) = \lambda\left|g(\vec{x})p(\vec{x})\right|$$
- If the function g(x)p(x) varied with x as the solid line in figure, then d(x) would be proportional to the dotted line. The value of λ may be found by requiring that ∫d(x)dx=1.



- If g(X)≥0, then d(x)= λg(x)p(x) and λ=1/I, so

$$d(\vec{x}) = \frac{g(\vec{x})p(\vec{x})}{I}$$

- A Monte Carlo algorithm to evaluate the integral would be to sample a series of x_i from d(x) and construct the sum

$$G_N = \frac{1}{N}\sum_{i=1}^{N} \frac{g(\vec{x}_i)p(\vec{x}_i)}{d(\vec{x}_i)} = \frac{1}{N}\sum_{i=1}^{N} I\frac{g(\vec{x}_i)p(\vec{x}_i)}{g(\vec{x}_i)p(\vec{x}_i)} = \frac{1}{N}\sum_{i=1}^{N} I = I$$

- If we already know the correct answer I, the Monte Carlo calculation will certainly give it back with zero variance! This clearly corresponds to the minimum variance calculation.

- Although we cannot in practice use this probability distribution, we expect that "similar" functions will reduce the variance.
- What is meant will be explored in some examples.

Some examples

- Consider the following integral to be computed via statistical sampling $I = \int_0^1 e^x dx = e^x\Big|_0^1 = e - 1$
- As a first Monte Carlo algorithm we can sample x uniformly on (0,1):

$$I = \int_0^1 e^x dx = \int_0^1 g(x)p(x)dx \quad with \quad p(x) = 1\,, \quad g(x) = e^x$$

it follows that the variance connected with this choice is

$$\sigma_I^2 = \left\langle g^2 \right\rangle_p - \left\langle g \right\rangle_p^2 = \int_0^1 e^{2x} dx - (e-1)^2 = \frac{1}{2}(e^2 - 1) - (e-1)^2 \approx 0.242$$

- Our second choice will be a probability density more similar to the integrand. Let's consider the Taylor's expansion of the exponential: exp(x)=1+x+... we can choose p(x)∝(1+x)

$$\int_0^1 (1+x)dx = \left(x + \frac{x^2}{2}\right)\Big|_0^1 = \frac{3}{2} \quad \Rightarrow \quad p(x) = \frac{2}{3}(1+x)$$

$$I = \int_0^1 e^x dx = \int_0^1 g(x)p(x)dx \quad with \quad p(x) = \frac{2}{3}(1+x)\,, \quad g(x) = \frac{3}{2}\frac{e^x}{1+x}$$

- the variance connected with this choice is

$$\sigma_I^2 = \langle g^2 \rangle_p - \langle g \rangle_p^2 = \int_0^1 \frac{9}{4}\frac{e^{2x}}{(1+x)^2}\frac{2}{3}(1+x)dx - (e-1)^2 =$$

$$= \frac{3}{2}\int_0^1 \frac{e^{2x}}{1+x}dx - (e-1)^2 \approx 0.027$$

which is about 10 times smaller with respect to the previous choice. Thus a Monte Carlo calculation of the previous integral with the choice p(x)=2(1+x)/3 for every fixed number of Monte Carlo steps, N_{MC} , will give an estimation of the integral with an uncertainty more than 3 times smaller (√10) than a uniform sampling.

- When the integrand g(x)p(x) of an integral is singular, var[g] may not exist. In this case we can always choose d(x) such that the ratio g(x)p(x)/d(x) is bounded. We assume here that g(x) and p(x) are known analytical functions whose singularities are easily identified. Consider, for example, the integral

$$I = \int_0^1 x^{-\frac{1}{2}}dx$$

the straightforward Monte Carlo calculation would be to sample **x** uniformly on [0,1] with the estimator $g=x^{-1/2}$. The variance contains $\langle g^2 \rangle$, which is

$$\langle g^2 \rangle = \int_0^1 x^{-1}\,dx = \infty$$

so that the variance for this calculation does not exist.

- As an alternative we can try with r<1.

$$d(x) = \frac{1-r}{x^r}$$

- The estimator for I is now

$$I = \int_0^1 \frac{dx}{\sqrt{x}} = \int_0^1 \frac{x^r}{(1-r)\sqrt{x}}\frac{1-r}{x^r}dx = \int_0^1 \frac{x^{r-\frac{1}{2}}}{(1-r)}\frac{1-r}{x^r}dx \quad \Rightarrow \quad g'(x) = \frac{x^{r-\frac{1}{2}}}{(1-r)}$$

and the n^{th} moment of g′ is

$$\langle g'^n \rangle = (1-r)^{1-n}\int_0^1 \frac{x^{nr-n/2}}{x^r}dx = (1-r)^{1-n}\int_0^1 x^{(n-1)r-n/2}dx$$

For this integral to exist, (n-1)r-n/2 > -1 , and in particular, all moments will exists if 1>r≥1/2. Of course, the optimal is r=1/2.

- Importance sampling is a very useful technique for reducing variance and for handling singular integrals. It will be used repeatedly in the applications of Monte Carlo methods to be discussed.

- A shortcoming of importance sampling is that it is difficult to apply if the integrand changes sign… the dawning of the sign problem. See more variance reduction techniques in the supplementary material.

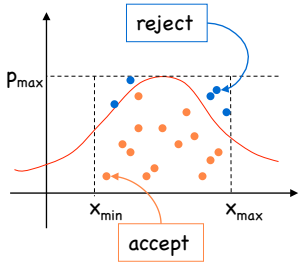
Rejection techniques

- One can compose random variables in a way that leads to very general techniques for sampling any probability distribution: a trial value for a random variable is selected and proposed. This value is subjected to one or more tests (involving one or more other random variables) and it may be accepted, that is, used as needed, or rejected. If it is rejected, the cycle of choosing and testing a trial value is repeated until an acceptance takes place. An important property of the method is that the normalization of the density and the cumulative distribution function need not be known explicitly to carry out the sampling.
- A disadvantage is that it may have low efficiency, that is, many values are rejected before one is accepted.
- Moreover it only works for probability distributions defined over a finite range.
- These drawbacks are balanced by the generality of the method, and certainly there are plenty of situations where the rejection method is the method of choice.

- In its simplest form, the rejection method works like this. We want to generate random numbers x in the interval from x_{min} to x_{max} distributed according to some function p(x). Let p_{max} be the maximum value which the function attains in the interval
  - We generate a random number x uniformly in the interval between x_{min} and x_{max} (This only works if x_{min} and x_{max} are finite, which is the reason for the comment above)
  - Now we generate another random number r between 0 and 1 and we keep the random number x if

$$r < \frac{p(x)}{p_{\max}}$$

    Otherwise, we reject x and generate another number between x_{min} and x_{max}



  - The process continues until we accept one of our numbers, and that is the number which our random number generator returns.
- The factor of p_{max} on the bottom of the previous equation ensures that the acceptance probability $p(x)/p_{max}$ has a maximum value of one, which makes the algorithm as efficient as possible. Why does it work?

- Well, the probability that our algorithm which samples uniformly the interval [x_{min} ,x_{max}] generates a number in some small range x to x+dx is

$$p_{gen}(x)dx = \frac{dx}{x_{\max} - x_{\min}}$$

- The probability that it is accepted comes from the equation in the previous slide. It is just:

$$p_{accept}(x) = \frac{p(x)}{p_{\max}}$$

- And the probability that we return a value of x in this interval is then the product of these two probabilities:

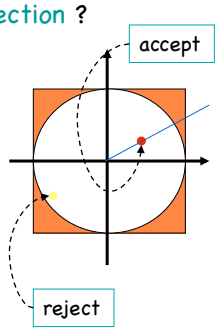
$$p_{gen}(x)p_{accept}dx = \frac{p(x)dx}{p_{\max}(x_{\max} - x_{\min})}$$

- Clearly this is proportional to p(x)dx, as we want it to be. Notice that there is no need in this case for the function p(x) to be normalized to unity over the range of allowed values for x. The factor of p_{max} ensures that the results of the algorithm will be unchanged if we multiply p(x) by any constant factor.

- The inefficiency of the algorithm arises for two reasons:
  - First, every time we try a new value of x we have to generate two new random numbers between zero and one. One of them is used to produce the value of x , and the other is used to decide whether to accept or reject that value.
  - Second, the very fact that we reject some candidate numbers reduces the algorithm's efficiency. It means that we waste time generating numbers and then throwing them away only to generate more.
- How to sample uniformly an angle in [0,2π] with rejection ?
  - Sample two uniform random numbers x,y in [-1,1] and check if $x^2+y^2<1$
  - If NO reject
  - If YES →

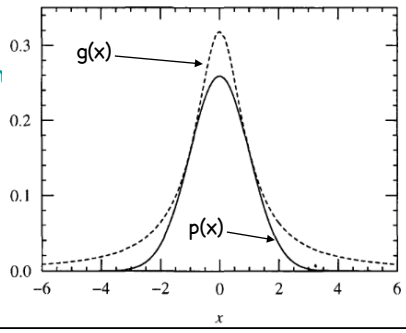
$$\begin{cases} \theta = \cos^{-1}\left(\frac{x}{\sqrt{x^2+y^2}}\right) & if \quad y \geq 0 \\ \theta = 2\pi - \cos^{-1}\left(\frac{x}{\sqrt{x^2+y^2}}\right) & if \quad y < 0 \end{cases}$$

  - In fact, uniform random number couples will fill uniformly the square, thus the points fallen inside the circle will fill it uniformly and the lines which connects the origin with these points individuate a random angle in [0,2π]



The hybrid method

- The rejection method described above is a simple alternative to the transformation method when we want to generate non-uniform random numbers with non-integrable distribution functions. As we saw, however, it is an inefficient method, requiring several calls to our basic random number generator for each number it returns.
- At the expense of a certain increase in programming complexity, we can reduce this inefficiency by using a hybrid method which combines features of both the transformation and rejection methods, while still being applicable to any distribution function p(x).
- The method works like this:
- First we must find a function g(x) which approximates to p(x), but which is integrable, and which is greater than or equal to p(x) for all x in the range between x_{min} and x_{max} .
- Then we apply the transformation method to generate a random number distributed according to g(x).



- Assuming g(x) is normalized, the probability of generating a number between x and x + dx is

$$p_{gen}(x)dx = g(x)dx$$

- Now we accept or reject this number in a manner similar to the rejection method, but with acceptance probability

$$p_{accept}(x) = \frac{p(x)}{g(x)}$$

- Thus the total probability of generating a number in the given interval is

$$p_{gen}(x)p_{accept}dx = \frac{p(x)}{g(x)}g(x)dx = p(x)dx$$

as it should be.

- The normal rejection method could be regarded as a special case of this algorithm in which g(x) is just equal to p_{max} for all x.
- Clearly however, the better g(x) approximates to f(x) the more efficient the algorithm will become.
- Note that p(x) does not have to be normalized to unity in order for the method to work. In fact, it cannot be normalized to unity, since we know that g(x) is normalized, and p(x) < g(x) for all x.
- The hybrid method also does not require that the limits x_{min} and x_{max} be finite.

Stochastic processes

- Now we want to go on to time-dependent phenomena and information unfolding with time. This leads to the concept of stochastic process:
- Definition: A stochastic process is a family $(X_t)_{t\in I}$ of random variables.
- The index set can either be a discrete set, for instance, I=ℕ, which would lead to a discrete time stochastic process, or it can be continuous, for instance, I=ℝ, which would lead to a continuous time stochastic process.
- For every fixed n (discrete time) or t (continuous time) X_n or X_t , respectively, is a stochastic variable and we can use the concepts and properties discussed previously

n-point joint probability

- First of all a stochastic process will be characterized by a n-point (joint) distribution function

$$p_n(x_1,t_1;\cdots;x_n,t_n)$$

- With the help of these n-point distribution functions, we can calculate time dependent moments and correlation functions of the stochastic process under study:

$$\langle X(t_1)\cdots X(t_n)\rangle = \int_{\mathbb{R}^n} dx_1\cdots dx_n\; x_1\cdots x_n\; p_n(x_1,t_1;\cdots;x_n,t_n)$$

- We can define a time dependent covariance matrix for two stochastic processes

$$\mathrm{cov}\left[X_i(t_1),X_j(t_2)\right] = \sigma_{ij}^2(t_1,t_2) = \left\langle X_i(t_1)X_j(t_2)\right\rangle - \left\langle X_i(t_1)\right\rangle\left\langle X_j(t_2)\right\rangle$$

the diagonal elements of this matrix, when properly normalized, are called autocorrelation functions and those off-diagonal are the cross-correlations functions.

joint probability properties

- The hierarchy of distribution functions p_n characterizes a stochastic process completely and fulfils the following three relations:

» Positivity:

$$p_n(x_1,t_1;\cdots;x_n,t_n) \geq 0$$

» Completeness:

$$\int p_n(x_1,t_1;\cdots;x_n,t_n)dx_n = p_{n-1}(x_1,t_1;\cdots;x_{n-1},t_{n-1})$$

» Probability measure:

$$\int p_1(x,t)dx = 1$$

- Definition: A stochastic process is called stationary iff for all n we have

$$p_n(x_1,t_1+\Delta t;\cdots;x_n,t_n+\Delta t) = p_n(x_1,t_1;\cdots;x_n,t_n)$$

Conditional probabilities

- The conditional probabilities we defined previously can be used to yield information about sub-ensembles of the stochastic process:

  we define the conditional probability for the events at (k+1) to (k+l) given the events 1 to k to be:

$$p_{k+l}(x_1,t_1;\cdots;x_{k+l},t_{k+l}) =$$

$$= p_{l|k}(x_{k+1},t_{k+1};\cdots;x_{k+l},t_{k+l}\,|\,x_1,t_1;\cdots;x_k,t_k) \times p_k(x_1,t_1;\cdots;x_k,t_k)$$

  Specifically we have

$$p_2(x_1,t_1;x_2,t_2) = p_{1|1}(x_2,t_2|x_1,t_1)p_1(x_1,t_1)$$

  and

$$\int p_{1|1}(x_2,t_2;x_1,t_1)dx_2 = 1$$

Markov processes

- Definition: For a Markov process, we have for all n and all $t_1<t_2<\ldots<t_n$

$$p_{1|n-1}(x_n,t_n|x_1,t_1;\cdots;x_{n-1},t_{n-1}) = p_{1|1}(x_n,t_n|x_{n-1},t_{n-1})$$

One therefore only has to know the actual state (x_{n-1},t_{n-1}) in order to calculate the probability for the occurrence of (x_n,t_n). In this sense, this process has no memory.

- If we apply this iteratively to the n-point function (joint probability distribution), we get

$$p_n(x_1,t_1;\cdots;x_n,t_n) = \prod_{l=2}^{n} p_{1|1}(x_l,t_l|x_{l-1},t_{l-1}) p_1(x_1,t_1)$$

In the case of a Markov process, therefore, the knowledge of two functions $p_{1|1}$ and p_1, suffices to describe the process completely

- An enormous simplification compared to the infinite hierarchy of p_n needed in the more general case (the "ideal gas" of the stochastic processes)

Chapman-Kolmogorov equation

- Let us now specialize to the case n=3

$$p_3(x_1,t_1;x_2,t_2;x_3,t_3) = p_{1|1}(x_3,t_3|x_2,t_2) p_{1|1}(x_2,t_2|x_1,t_1) p_1(x_1,t_1)$$

integrating over x_2 and using the completeness property of the hierarchy of the p_n we get

$$p_2(x_1,t_1;x_3,t_3) = \int p_3(x_1,t_1;x_2,t_2;x_3,t_3)\,dx_2 = p_1(x_1,t_1)\int p_{1|1}(x_3,t_3|x_2,t_2) p_{1|1}(x_2,t_2|x_1,t_1)\,dx_2$$

with the definition of the conditional probability, we arrive at the following result

$$p_{1|1}(x_3,t_3|x_1,t_1) = \int p_{1|1}(x_3,t_3|x_2,t_2) p_{1|1}(x_2,t_2|x_1,t_1)\,dx_2 \quad for \quad t_3 \geq t_2 \geq t_1$$

This is the Chapman-Kolmogorov equation. It is a consistency equation for the conditional probabilities of a Markov process.

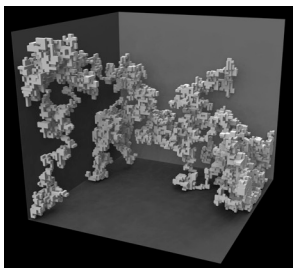
- Theorem: Two positive, normalized functions p_1 and $p_{1|1}$ which fulfil
  - The Chapman-Kolmogorov equation
  - The equation: $p_1(x_2,t_2) = \int p_{1|1}(x_2,t_2|x_1,t_1) p_1(x_1,t_1)\,dx_1$

completely and uniquely define a Markov process

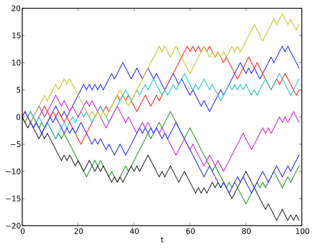
Random walk on a lattice

- A random walk is a mathematical formalisation of a trajectory that consists of taking successive random steps.
- In simple random walk, the walk can only jump to neighbouring sites of a regular lattice according to some probability distribution.
- The most well-studied example is of random walk on the *d*-dimensional integer lattice $\mathbb{Z}^d$
- Model:
  (1) a particle can be found in one of these lattice sites
  (2) at each (discrete) instant the particle make a jump moving towards one of the nearest neighbour lattice sites

3D random walk



1D random walks



1D Random walk

- The problem of random walk on a lattice can be formulated as a Markov process: let us consider random walk on an infinite 1D lattice with lattice spacing, a, along the x axis, and let us assume that the time between steps is τ.
- Let $p_1(na,s\tau)$ be the probability to find the particle at point x = na ($n \in \mathbb{Z}$) after s steps. Then

$$p_1(na,(s+1)\tau) = \sum_{m=-\infty}^{\infty} p_{1|1}(na,(s+1)\tau|ma,s\tau)\,p_1(ma,s\tau)$$

  where $p_{1|1}$ in the previous equation is the transition probability to go from site x = ma to site x = na in one step.
- As a specific example, let us consider the case in which the random walker has an equal probability to go one lattice site to the left or right during each step.

Continuum limit

- In this case, the transition probability is

$$p_{1|1}\left(na,(s+1)\tau \middle| ma,s\tau\right) = \tfrac{1}{2}\delta_{n,m+1} + \tfrac{1}{2}\delta_{n,m-1}$$

and thus

$$p_1\left(na,(s+1)\tau\right) = \tfrac{1}{2}p_1\left((n+1)a,s\tau\right) + \tfrac{1}{2}p_1\left((n-1)a,s\tau\right)$$

- We can obtain a differential equation for the probability $p_1(na,s\tau)$ in the continuum limit

- Let us subtract $p_1(na,s\tau)$ from both sides and divide by τ:

$$\frac{p_1\left(na,(s+1)\tau\right) - p_1\left(na,s\tau\right)}{\tau} = \frac{a^2}{2\tau}\left[\frac{p_1\left((n+1)a,s\tau\right) + p_1\left((n-1)a,s\tau\right) - 2p_1\left(na,s\tau\right)}{a^2}\right]$$

- If we now let $x = na$, $t = s\tau$, and take the limit $a\to 0$ and $\tau\to 0$ so that $D=a^2/2\tau$ is finite and x and t are finite …

A diffusion process

- … we obtain the following differential equation for p_1:

$$\frac{\partial p_1(x,t)}{\partial t} = D\frac{\partial^2 p_1(x,t)}{\partial x^2}$$

which is a diffusion equation for the probability density p_1

- Let us solve for the case $p_1(x,t=0)=\delta(x)$. We first introduce the Fourier transform of $p_1(x,t)$

$$\hat{p}_1\left(k,t\right) = \int_{-\infty}^{\infty} dx\; e^{-ikx} p_1\left(x,t\right)$$

then the previous equation takes the form

$$\frac{\partial \hat{p}_1(k,t)}{\partial t} = -Dk^2\hat{p}_1(k,t) \quad\Rightarrow\quad \hat{p}_1(k,t) = e^{-Dk^2t}$$

where we have used that $\hat{p}_1(k,0) = 1 \quad \left(p_1(x,t=0) = \delta(x)\right)$

it follows that

$$p_1\left(x,t\right) = \tfrac{1}{2\pi}\int_{-\infty}^{\infty} dk\; e^{ikx}e^{-Dk^2t} = \tfrac{1}{\sqrt{4\pi Dt}}e^{-\frac{x^2}{4Dt}}$$

and easily that $\left\langle x(t)\right\rangle = 0 \quad ; \quad \left\langle x^2(t)\right\rangle = 2Dt$

CLT and random walks

- Note that the "sum" variable, $S_N=\Sigma_i x_i$, introduced in the CLT when

$$p(x) = 0.5\left[\delta(x+a) + \delta(x-a)\right]$$

is nothing else than a realization of a 1D random walk on a lattice

- We have in this case:

$$\langle x \rangle = 0 \qquad \sigma^2 = \int x^2 p(x) dx = a^2 \qquad \Rightarrow \sigma^2_{S_N} = Na^2$$

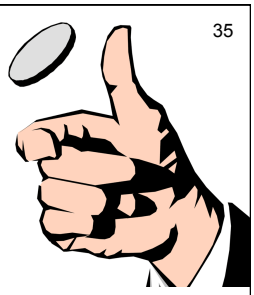
- It follows that the CLT limiting distribution is:

$$P_N(S_N) \underset{N\to\infty}{\to} \frac{1}{\sqrt{2\pi N a^2}} \exp\left[-\frac{S_N^2}{2Na^2}\right]$$

which is consistent with the previous result. In fact, given that the time elapsed is $t=N\tau$, the identification $Na^2=2Dt=2DN\tau$ turns out in the relation: $D=a^2/2\tau$

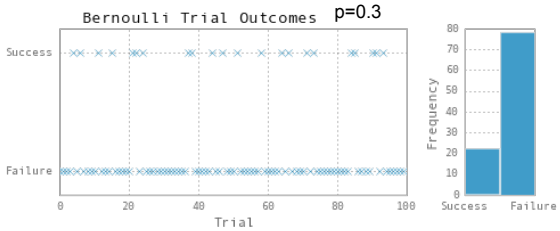
- We have noticed the connection between the limiting distribution of the CLT and the limiting distribution of a random walk; moreover, we have observed that this distribution fulfils a partial differential eq.

Bernulli process



- The simplest discrete stochastic process is the Bernulli process, which can be seen as a repeated coin flipping, possibly with an unfair coin (but with consistent unfairness)
- A Bernoulli process is a sequence of independent random variables x_1, x_2, x_3, …, such that:

1. for each i, the value of x_i is either 0 or 1 (yes/no, success/failure, etc.)
2. For all values of i, the probability that $x_i=1$ is the same number p



- Independence of the trials implies that the process is memoryless. A single trial for a Bernoulli process, is called a Bernoulli trial.

- Starting from a Bernulli process we can construct other different stochastic processes. For example if you ask the probability to have k success among n Bernoulli trials, then the answer will have a binomial distribution, B(n,p); formally, it's the sum $B_n=x_1+x_2+\cdots+x_n$ of n i.i.d. Bernoulli trials

$$\Pr(B_n = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

- The formula can be understood as follows: k successes occur with probability p^k and n-k failures occur with probability $(1-p)^{n-k}$. However, the k successes can occur anywhere among the n trials, and there are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ different ways of distributing k successes in a sequence of n trials.

- This discrete stochastic markov process is characterized by the following finite difference eq.: $B_n = B_{n-1} + x_n \qquad x_n = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1-p \end{cases}$

- One can easily construct also a 1D random walk (RW) on a lattice with lattice constant a:

$$RW_n = RW_{n-1} + 2a(x_n - 0.5) \qquad x_n = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1-p \end{cases}$$

- Given that the sum $B_n=x_1+x_2+\cdots+x_n$ of n i.i.d. Bernoulli trials is distributed like

$$\Pr(B_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

with: $\langle B_n \rangle = np \qquad \sigma^2_{B_n} = np(1-p)$

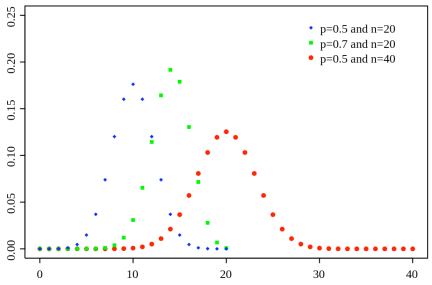
the CLT applies and the binomial distribution must converge to a Gaussian for n→∞: $\mathcal{N}(np,np(1-p))$

$$P_n(B_n) \underset{N\to\infty}{\to} \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left[-\frac{(B_n - np)^2}{2np(1-p)}\right]$$

- de Moivre–Laplace theorem: As n approaches ∞ while p remains fixed, the distribution of

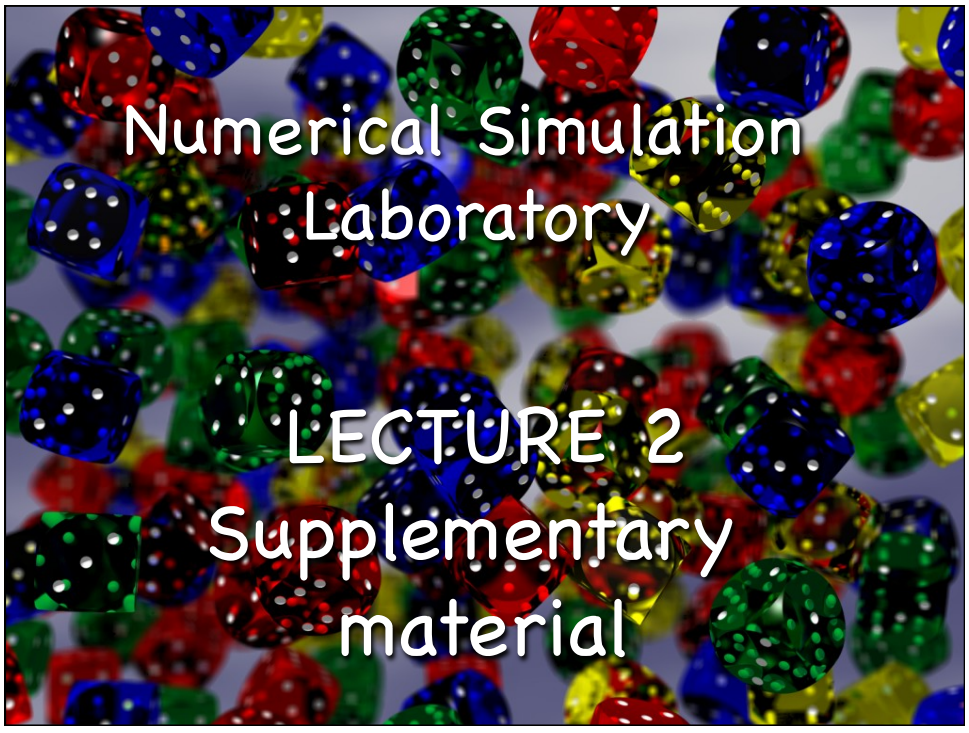
$$\frac{B_n - np}{\sqrt{np(1-p)}}$$

approaches the normal distribution with mean=0 and variance=1: $\mathcal{N}(0,1)$



Lecture 2: Suggested books

- Kalos & Whitlock, *Monte Carlo Methods* - Wiley 1986
- Bratley, Fox, Schrage, *A Guide to Simulation* - Springer 1987
- W. Paul & J. Baschnagel, *Stochastic Processes* – Springer 2013
- C.W. Gardiner, *Handbook of Stochastic Methods* – Springer 2004
- E. Vitali, M. Motta, D.E. Galli, *Theory and Simulation of Random Phenomena* – Springer 2018



Variance reduction techniques

- Three major classes of techniques are used to reduce the variance in Monte Carlo quadrature:

1. Importance sampling can be introduced into the calculation to increase the likelihood of sampling variables where the function is large or rapidly varying.
2. The expected value of a random variable can be used rather than the variable itself. This substitution never increases variance and many times will substantially reduce it.
3. Correlations between succeeding samples may be exploited to advantage. In control variates, an easily evaluated approximation to the integrand is used to reduce the variance. If successive random variables are negatively correlated, the variance will be smaller than if they were independent. The technique called antithetic variates exploits the reduction in variance that results when negatively correlated samples are deliberately produced and grouped together.

The use of expected values…

- The following discussion describes the use of expected values in finite quadrature to reduce the variance but its application is much more wide ranging. Suppose we wish to evaluate the integral

$$I = \int g(x,y)p(x,y)dxdy$$

x and y may be many-dimensional vectors (though y is usually 1D).

- The marginal distribution for X is defined by $m(x) = \int p(x,y)dy$ and we can define another quantity h(x) as:

$$h(x) = \frac{1}{m(x)} \int g(x,y)p(x,y)dy$$

- We assume that the integrals m(x) and h(x) can be evaluated by means other than Monte Carlo. The integral I can be rewritten as

$$I = \int h(x)m(x)dx$$

We assume that the order of integration is immaterial.

- One can easily prove that $\mathrm{var}[g]_p - \mathrm{var}[h]_m \geq 0$
- In other words, the variance of a Monte Carlo calculation may be reduced by doing part of the integration analytically

Control variates

- Correlation methods serve to reduce the variance by the use of correlated points in the sampling rather than sampling all points independently.
- In a technique called control variates, the integral of interest,

$$I = \int g(x)p(x)dx$$

is written as $I = \int [g(x) - h(x)]p(x)dx + \int h(x)p(x)dx$
where ∫h(x)p(x)dx is known analytically.

- The estimator for I becomes

$$I \cong \int h(x)p(x)dx + \frac{1}{N}\sum_{i=1}^{N}[g(x_i) - h(x_i)]$$

with g(x) and h(x) evaluated at the same points x_i .

- The technique is advantageous when $\mathrm{var}[g-h]_p << \mathrm{var}[g]_p$ and this occurs when h(x) is very similar to g(x).
- If ∫g(x)p(x)dx closely resembles a known integral, then the method will probably be useful.

- Consider the integral discussed earlier

$$I = \int_0^1 e^x dx$$

we have seen that the variance associated with a straightforward Monte Carlo evaluation is about 0.242

- A possible h(x) derives from the first two terms in the Taylor series of e^x ,

$$\int_0^1 e^x dx = \int_0^1 [e^x - (1+x)]dx + \tfrac{3}{2}$$

- The random variable x may be chosen uniformly on [0,1] and the associated variance is

$$\sigma_I^2 = \int_0^1 [e^x - (1+x)]^2 dx - [e - 1 - \tfrac{3}{2}]^2 \cong \int_0^1 (e^{2x} + 1 + x^2 + 2x - 2e^x - 2xe^x)dx - 0.0476 =$$

$$= \left[\frac{e^{2x}}{2} + x + \frac{x^3}{3} + x^2 - 2e^x - 2(xe^x - e^x)\right]_0^1 - 0.0476 = \frac{e^2}{2} + \frac{7}{3} - 2e - \frac{1}{2} - 0.0476 =$$

$$\cong 0.0913 - 0.0476 = 0.0437$$

which is a substantial reduction from the variance quoted above.

Antithetic variates

- The method of antithetic variates exploits the decrease in variance that occurs when random variables are negatively correlated. When variables are negatively correlated, if the first point gives a value of the integrand that is larger than average, the next point will be likely to give a value that is smaller than average, and the average of the two values will be closer to the actual mean. In theory the method sounds promising, but in practice it has not been very successful as a general variance reduction method in many dimensions.
- Suppose for example that $I = \int_0^1 g(x)dx$ where g(x) is linear.
  Then I may be written exactly as $I = \int_0^1 \frac{1}{2}\left[g(x) + g(1-x)\right]dx$
- I may be evaluated through Monte Carlo by using the last equation and picking an x uniformly on [0,1]. The value of g(x) and g(1-x) is determined and the estimate for I formed,

$$G_N = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\left[g(x_i) + g(1-x_i)\right]$$

  which will give exactly I with zero variance for linear g(x).
- For nearly linear functions, this method will substantially reduce the variance.

- Consider the integral discussed in the previous lecture

$$I = \int_0^1 e^x dx$$

  we have seen that the variance associated with a straightforward Monte Carlo evaluation is about 0.242
- If we pick $x_1, x_2, \ldots, x_N$ uniformly and at random on [0,1] and form the estimator

$$G_N = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\left[g(x_i) + g(1-x_i)\right] = \frac{1}{N}\sum_{i=1}^{N} \frac{e^{x_i} - e^{1-x_i}}{2}$$

  the variance associated with this calculation is

$$\sigma_I^2 = \left\langle \left[\frac{g(x)+g(1-x)}{2}\right]^2 \right\rangle_{uniformly} - I^2 = \frac{1}{4}\int_0^1 \left(e^{2x} + e^{2-2x} + 2e\right)dx - (e-1)^2 =$$

$$= \frac{1}{4}\left(\frac{e^{2x}}{2} - \frac{e^{2-2x}}{2} + 2ex\right)\Bigg|_0^1 - (e-1)^2 = \frac{1}{4}\left(\frac{e^2}{2} - \frac{1}{2} + 2e\right) - \frac{1}{4}\left(\frac{1}{2} - \frac{e^2}{2}\right) - (e-1)^2 =$$

$$= \frac{e^2}{4} - e^2 + \frac{e}{2} + 2e - \frac{1}{4} - 1 = -\frac{3e^2}{4} + \frac{5e}{2} - \frac{5}{4} \cong 0.0039$$

which is a dramatic reduction in variance.

- Correlation techniques need not be used alone, but may be combined with other methods of reducing the variance. For example, both importance sampling and antithetic variables may be used simultaneously to improve a calculation. Thus the previous example may be improved by importance sampling as

$$I = \int_0^1 \frac{1}{2} \frac{g(x) + g(1-x)}{p(x)} p(x)dx$$

- The antithetic estimator $[e^x+e^{1-x}]/2$ is symmetric about x=1/2 , so an approximate p(x) for this problem is

$$p(x) = \frac{24}{25}\left[1 + \frac{1}{2}\left(x - \frac{1}{2}\right)^2\right]$$

  chosen to agree with three terms of the power series at x=1/2.
- The variance of the relative estimator with this sampling function is 0.0000012, compared with 0.0039 for the previous estimator and 0.242 for the straightforward evaluation.
- Thus, use of antithetic variates reduces the variance by two orders of magnitude, and this simple importance sampling by another three!