Computational Astrophysics I: Introduction and basic concepts

Helge Todt

Astrophysics
Institute of Physics and Astronomy
University of Potsdam

SoSe 2020
Monte-Carlo integration
Idea: Can the area of a pool (irregular!) be measured by throwing stones?

- pool with area $F_n$ in a field with area $A$
fraction of the *randomly* thrown stones which fall into the pool:

\[ \frac{n_p}{n} = \frac{F_n}{A} \]  \hspace{1cm} (1)

(n stones, \( n_p \) hit pool)

determine \( F_n \) with help of the *hit-or-miss method*:

\[ F_n = A \frac{n_p}{n} \]  \hspace{1cm} (2)
rectangle of height $h$, width $(b - a)$, area $A = h \cdot (b - a)$, such that $f(x)$ within the rectangle

- generate $n$ pairs of random variables $x_i, y_i$ with $a \leq x_i \leq b$ and $0 \leq y_i \leq h$

- fraction $n_t$ of the points, which fulfill $y_i \leq f(x_i)$ gives estimate for area under $f(x)$ (integral)
Buffon’s needle problem – determine $\pi$ by throwing matches

Buffon’s question (1773): What is the probability that a needle or a match of length $\ell$ will lie across a line between two strips on a floor made of parallel strips, each of same width $t$?

$\rightarrow x$ is distance from center of the needle to closest line, $\theta$ angle between needle and lines, hence the uniform probability density functions are

$$p(x) = \begin{cases} \frac{2}{t} & : \quad 0 \leq x \leq \frac{t}{2} \\ 0 & : \quad \text{elsewhere} \end{cases} \quad p(\theta) = \begin{cases} \frac{2}{\pi} & : \quad 0 \leq \theta \leq \frac{\pi}{2} \\ 0 & : \quad \text{elsewhere} \end{cases}$$

$x, \theta$ independent $\rightarrow p(x, \theta) = \frac{4}{t\pi}$ with condition $x \leq \frac{\ell}{2} \sin \theta$. If $\ell \leq t$ (short needle):

$$P(\text{hit}) = \int_{\theta=0}^{\frac{\pi}{2}} \int_{x=0}^{\frac{\ell}{2} \sin \theta} \frac{4}{t\pi} \, dx \, d\theta = \frac{2\ell}{t\pi}$$

$\rightarrow$ count hits and misses and then:

$$\pi = \frac{2\ell}{t} \frac{n_{\text{hit}} + n_{\text{miss}}}{n_{\text{hit}}}$$
Sample-mean method

- the integral

\[
F(x) = \int_a^b f(x) \, dx
\]  

(3)

is given in the interval \([a, b]\) by the mean \(\langle f(x) \rangle\) (mean value theorem for integration)

- choose arbitrary \(x_i\) (instead of regular intervals) and calculate

\[
F_n = (b - a)\langle f(x) \rangle = (b - a) \frac{1}{n} \sum_{i=1}^{n} f(x_i)
\]  

(4)

where \(x_i\) are uniform random numbers in \([a, b]\)

\[
\left( \text{cf. rectangle rule} \quad F_n = \sum_{i=1}^{n-1} f(x_i)\Delta x \quad \text{with fixed } x_i, \Delta x \right)
\]  

(5)
Error estimation
Numerical integration (exact or MC) gives approximation

\[ \int_{a}^{b} f(x) \, dx = Q(f) + E(f) \]  

(6)

\( Q(f) \) so-called quadrature formula,
\( E(f) \) error \( \rightarrow \) unknown (obvious)

Aim: estimate magnitude of error

so far: error calculated from our knowledge of the exact result
Obvious: for constant integrand $f$ is $E = 0$, i.e. $F_n$ is independent of $n$ (and always the same)

Idea: try to estimate the error with help of the standard deviation $\sigma$:

$$\sigma^2 = \langle f(x)^2 \rangle - \langle f(x) \rangle^2$$  \hspace{1cm} (7)

$$\langle f(x) \rangle = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$ \hspace{1cm} (8)

$$\langle f(x)^2 \rangle = \frac{1}{n} \sum_{i=1}^{n} f(x_i)^2$$ \hspace{1cm} (9)

if $f$ constant $\rightarrow \sigma = 0$
consider the example \( f(x) = 4\pi \sqrt{1 - x^2} \) with \( F = \int_0^1 f(x)\,dx = \pi \)

Calculate \( \sigma \) for different \( n \)

| \( F_n \)    | \( n \) | \( E = |F_n - \pi| \) | \( \sigma \) |
|-------------|--------|-------------------|--------|
| 3.271771    | 10\(^1\) | 0.13017           | 0.78091 |
| 3.100276    | 10\(^2\) | 0.04131           | 0.91441 |
| 3.173442    | 10\(^3\) | 0.03185           | 0.85013 |
| 3.135863    | 10\(^4\) | 0.00572           | 0.90317 |
| 3.142189    | 10\(^5\) | 0.00059           | 0.89051 |
| 3.141798    | 10\(^6\) | 0.00020           | 0.89236 |

- \( \sigma \) almost constant and much larger than \( E \)
- but: decrease of \( E \) from \( n = 10^2 \) to \( n = 10^4 \) by a factor of 10 → \( \sim 1/n^{1/2} \)
- therefore: \( \sigma \) says how much \( f \) varies in \([a, b]\)
Numerical integration and error IV

- idea: estimate $E$ by several runs $\alpha$ for constant $n = 10^4$, each with result $M_\alpha$:

<table>
<thead>
<tr>
<th>$M_\alpha$</th>
<th>$\alpha$</th>
<th>$E$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.14892</td>
<td>1</td>
<td>0.00735</td>
</tr>
<tr>
<td>3.13255</td>
<td>2</td>
<td>0.00904</td>
</tr>
<tr>
<td>3.14042</td>
<td>3</td>
<td>0.00117</td>
</tr>
<tr>
<td>3.14600</td>
<td>4</td>
<td>0.00441</td>
</tr>
<tr>
<td>3.15257</td>
<td>5</td>
<td>0.01098</td>
</tr>
<tr>
<td>3.13972</td>
<td>6</td>
<td>0.00187</td>
</tr>
<tr>
<td>3.13107</td>
<td>7</td>
<td>0.01052</td>
</tr>
<tr>
<td>3.13585</td>
<td>8</td>
<td>0.00574</td>
</tr>
<tr>
<td>3.13442</td>
<td>9</td>
<td>0.00717</td>
</tr>
<tr>
<td>3.14047</td>
<td>10</td>
<td>0.00112</td>
</tr>
</tbody>
</table>

- $E$ varies, differences $|M_\alpha - M_\beta|_{\alpha \neq \beta}$ between results comparable with $E$, therefore:
define standard deviation $\sigma_m$ of the means:

$$\sigma_m^2 = \langle M^2 \rangle - \langle M \rangle^2$$

(10)

$$\langle M \rangle = \frac{1}{m} \sum_{\alpha=1}^{m} M_\alpha \quad \rightarrow \quad \langle M^2 \rangle = \frac{1}{m} \sum_{\alpha=1}^{m} M_\alpha^2$$

(11)

(12)

for the runs 1 till 10 one gets $\sigma_m = 0.006762 \rightarrow$ comparable with $E$

- exact: one run has the chance of 68% that $M_\alpha$ is in in the range $\pi \pm \sigma_m$

- however method not very useful, as several runs are required
Numerical integration and error VI

actually for large $n$ holds:

$$
\sigma_m = \frac{\sigma}{\sqrt{n-1}} \approx \frac{\sigma}{\sqrt{n}}
$$

(13)

e.g., for $n = 10^4$ is $\sigma_m = 0.90317/100 \approx 0.009$, i.e., consistent with our estimate $\sigma_m = 0.007$ and the error $E = 0.006$

How can we get $\sigma$ without $\alpha$ runs?
Hence, split one run, e.g., in $s = 10$ subsets $k$ such that each contains $n/s = 1000$ trials and has result $S_k$

Then, with the mean $\langle S \rangle$ is also

$$\sigma_s^2 = \langle S^2 \rangle - \langle S \rangle^2$$ \hspace{1cm} (14)

and

$$\sigma_m = \sigma_s / \sqrt{s}$$ \hspace{1cm} (15)
derivation:

- random variable $x$
- $m$ runs with each $n$ trials ($= mn$ trials in total)
- index $\alpha$ lables a run, $i$ a single trial

result from one run ($= measurement$):

$$M_\alpha = \frac{1}{n} \sum_{i=1}^{n} x_{\alpha,i}$$  \hspace{1cm} (16)$$

the arithmetic mean of all $mn$ trials is:

$$\overline{M} = \frac{1}{m} \sum_{\alpha} M_\alpha = \frac{1}{nm} \sum_{\alpha=1}^{m} \sum_{i=1}^{n} x_{\alpha,i}$$  \hspace{1cm} (17)$$
difference of a one run $\alpha$ and the total mean

$$e_\alpha = M_\alpha - \overline{M}$$  \hspace{1cm} (18)

Hence the variance (standard deviation$^2$) can be written for the runs as:

$$\sigma^2_m = \frac{1}{m} \sum_{\alpha=1}^{m} e^2_\alpha$$  \hspace{1cm} (19)

Now finding the relation between $\sigma_m$ and $\sigma$ of the individual $mn$ trials. Difference between one trial and the mean of one run:

$$d_{\alpha,i} = x_{\alpha,i} - \overline{M}$$  \hspace{1cm} (20)

Therefore the variance for all $mn$ trials:

$$\sigma^2 = \frac{1}{mn} \sum_{\alpha=1}^{m} \sum_{i=1}^{n} d^2_{\alpha,i}$$  \hspace{1cm} (21)
With help of Eq. (20) the Eq. (18) can be rewritten as:

\[ e_\alpha = M_\alpha - \bar{M} = \frac{1}{n} \sum_{i=1}^{n} (x_{\alpha,i} - \bar{M}) = \frac{1}{n} \sum_{i=1}^{n} d_{\alpha,i} \]  

(22)

Insert Eq. (22) into Eq. (19):

\[ \sigma_m^2 = \frac{1}{m} \sum_{\alpha=1}^{m} \left( \frac{1}{n} \sum_{i=1}^{n} d_{\alpha,i} \right) \left( \frac{1}{n} \sum_{j=1}^{n} d_{\alpha,j} \right) \]  

(23)
The products in Eq. (23) consist of terms $i = j$ and terms $i \neq j$. As the trials are independent of each other, for large $n$ the differences $d_{\alpha,i}$ and $d_{\alpha,j}$ are on average as often negative as positive, i.e., the terms $i \neq j$ cancel out on average. What remains are the terms for $i = j$:

$$
\sigma_m^2 = \frac{1}{mn^2} \sum_{\alpha=1}^{m} \sum_{i=1}^{n} d_{\alpha,i}^2
$$

(24)

By comparison with Eq. (21) one gets:

$$
\sigma_m^2 = \frac{\sigma^2}{n} \Rightarrow \sigma_m = \frac{\sigma}{\sqrt{n}}
$$

(25)
Importance sampling

Idea: improve MC integration by a better sampling \(\rightarrow\) introduce a positive function \(p(x)\) with

\[
\int_a^b p(x) \, dx = 1 \tag{26}
\]

and rewrite integral \(\int_a^b f(x) \, dx\) as

\[
F = \int_a^b \left[ \frac{f(x)}{p(x)} \right] p(x) \, dx \tag{27}
\]

this integral can be evaluated by *sampling according to* \(p(x)\):

\[
F_n = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x)}{p(x)} \tag{28}
\]
Importance sampling II

Note that for the uniform case \( p(x) = 1/(b - a) \) the sample mean method is recovered.

Now, try to minimize variance \( \sigma^2 \) of integrand \( \frac{f(x)}{p(x)} \) by choosing 

\[ p(x) \approx f(x), \text{ especially for large } f(x) \]

\[ \rightarrow \text{slowly varying integrand } f(x)/p(x) \]

\[ \rightarrow \text{smaller variance } \sigma^2 \]
Example: Normal distribution

Evaluate integral \( F = \int_{a}^{b} f(x) \, dx = \int_{0}^{1} e^{-x^2} \, dx \) (error function)

<table>
<thead>
<tr>
<th>( p(x) = 1 )</th>
<th>( p(x) = Ae^{-x} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>( (b - a) \times r + a )</td>
</tr>
<tr>
<td>( n )</td>
<td>4 \times 10^5</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.0404</td>
</tr>
<tr>
<td>( \sigma / \sqrt{n} )</td>
<td>6 \times 10^{-5}</td>
</tr>
<tr>
<td>total CPU time( ^\dagger )</td>
<td>19 ms</td>
</tr>
<tr>
<td>CPU time / trial</td>
<td>50 ns</td>
</tr>
</tbody>
</table>

\( ^\dagger \)CPU time on a Intel Core i7-4771 3.5 GHz

→ the extra time needed per trial for getting \( x \) from uniform \( r \) is usually overcompensated by the smaller number of necessary trials for same \( \sigma / \sqrt{n} \)
Similar: **Metropolis algorithm** (Metropolis, Rosenbluth, Rosenbluth, Teller & Teller 1953)

useful for averages of the form

\[
\langle f \rangle = \frac{\int p(x)f(x)dx}{\int p(x)dx}
\]  

(29)

Metropolis algorithm produces *random walk* (see below) of points \( \{x_i\} \) (1D or higher) with asymptotic probability distribution approaching \( p(x) \) for \( n \gg 1 \). Random walk from *transition probability* \( T(x_i \rightarrow x_j) \), such that

\[
p(x_i)T(x_i \rightarrow x_j) = p(x_j)T(x_j \rightarrow x_i) \quad \text{(detailed balance)}
\]  

(30)

e.g., \( T(x_i \rightarrow x_j) = \min \left[ 1, \frac{p(x_j)}{p(x_i)} \right] \)  

(31)
### Metropolis algorithm

1. Choose trial position $x_{\text{trial}} = x_i + \delta_i$ with random $\delta_i \in [-\delta, +\delta]$.
2. Calculate $w = p(x_{\text{trial}})/p(x_i)$.
3. If $w \geq 1$, accept and $x_{i+1} = x_{\text{trial}}$.
4. If $w < 1$, generate random $r \in [0; 1]$.
5. If $r \leq w$, accept and $x_{i+1} = x_{\text{trial}}$.
6. If not, $x_{i+1} = x_i$.

→ Problem: optimum choice of $\delta$; if too large, only small number of accepted trials → inefficient sampling.

If too small, only slow sampling of $p(x)$. Hence, rule of thumb: choose $\delta$ for which $\frac{1}{3} \ldots \frac{1}{2}$ trials accepted.

Also: choose $x_0$ for which $p(x_0)$ is largest → faster approach of $\{x_i\}$ to $p(x)$.
Typical applications for Metropolis algorithm: computation of integrals with weight functions $p(x) \sim e^{-x}$, e.g.,

$$\langle x \rangle = \frac{\int_0^\infty x e^{-x} \, dx}{\int_0^\infty e^{-x} \, dx} \quad (32)$$

$$\langle A \rangle = \frac{\int A(\vec{X}) e^{-U(\vec{X})/k_B T} \, d\vec{X}}{\int e^{-U(\vec{X})/k_B T} \, d\vec{X}} \quad (33)$$

where the latter is the average of a physical quantity $A$ in a liquid system with good contact to a thermal bath, fixed number of particles (with $\vec{X} = (\vec{x}_1, \vec{x}_2, \ldots)$ of all particles) and volume $\rightarrow$ canonical ensemble, e.g.,

$$\left\langle \frac{mv_{ik}^2}{2} \right\rangle = \frac{1}{2} k_B T \quad (34)$$
Rejection sampling
(acceptance-rejection method)
Rejection sampling (acceptance-rejection method)

Problem: get random $x$ for any $p(x)$, also if $P(r)^{-1}$ not (easily) computable

Idea:

- area under $p(x)$ in $[x, x + dx]$ is probability of getting $x$ in that range
- if we can choose a random point in *two dimensions* with uniform probability in the area under $p(x)$, then $x$ component of that point is distributed according to $p(x)$
- so, on same graph draw $f(x)$ with $f(x) > p(x) \ \forall x$
- if we can uniformly distribute points in the area under curve $f(x)$, then all points $(x, y)$ with $y < p(x)$ are uniform under $p(x)$
creation of arbitrary probability distributions with help of rejection sampling (especially for compact intervals $[a, b]$):

- let $p(x)$ be the required distribution in $[a, b]$
- choose a $f(x)$ such that $p(x) < f(x)$ in $[a, b]$, e.g., $f(x) = c \cdot \max(p(x)) = \text{const.}$ where $c > 1$
- it is $A := \int_a^b f(x)dx$, i.e. $A(x)$ must exist and be invertible
- generate uniform random number in $[0, A]$ and get the corresponding $x(A)$
- generate 2nd uniform random number $y$ in $[0, f(x)]$, so $x, y$ are uniformly distributed on $A$ (area under $f(x)$)
- accept this point if $y < p(x)$, otherwise reject it
Example: normal distribution $p(x)$ sampled by $f(x) = (x^2 + 1)^{-1}$

$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ (blue solid line) sampled with help of the function $\frac{1}{x^2+1}$ (red dashed) whose integral is $\arctan(x)$ (thick dashed red) and hence $F(x)^{-1} = \tan(x)$.
Requirements:

- $p(x)$ must be computable for every $x$ in the interval $[a, b]$.
- $f(x) > p(x)$ always possible, as $\int_{-\infty}^{+\infty} p(x)dx = 1$ (i.e. $A > 1$).
- to get $x_0$ for a chosen value in $[0, A]$ requires usually: $\int f(x)dx = F$ is analytically invertible, i.e. $F(x)^{-1}$ exists.

An example is easy for compact $[a, b]$, e.g.,

$$F(x) = c \cdot \max(p(x)) \cdot (x - a) = k(x - a) \rightarrow x = F/k - a$$

for randomly chosen $F$ in $[0, A]$, where $A = k \cdot (b - a)$. 
Example: acceptance-rejection for normal distribution

double p(double x){ return exp(-0.5*x*x)/sqrt(2.*M_PI); }
double f(double x){ return 1./(x*x+1.); }
double inv_int_f(double ax){ return tan(ax - M_PI /2.); }
...
for (int i = 0; i < nmax; ++i){
  // get random value between 0 and A:
  ax = A * double(rand())/double(RAND_MAX);
  // obtain the corresponding x value:
  x = inv_int_f(ax);
  // get random y value in interval [0,f(x)]:
  y = f(x) * double(rand())/double(RAND_MAX);
  // test for y <= p(x) for acceptance:
  if ( y <= p(x) ) { cout << x << endl ;}
In our example:

- it is \( p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \) the standard normal distribution; normal distributions with \( \sigma \neq 1, \mu \neq 0 \) can be obtained by transformation
- the comparison function \( f(x) = \frac{1}{x^2 + 1} \) is always \( f(x) > p(x) \), moreover:
  - \( F(x) = \int_{-\infty}^{x} f(x')dx' = \arctan(x) - \arctan(-\infty) = \arctan(x) - (-\frac{\pi}{2}) \)
    \( \rightarrow F(x) = \arctan(x) + \frac{\pi}{2} \)
  - the total area \( A \) under \( f(x) \) is
    \[ \int_{-\infty}^{\infty} f(x')dx' = \arctan(+\infty) - \arctan(-\infty) = \pi \]
  - the inverse \( F(x)^{-1} \), which returns \( x \) for a given value \( F \in [0, A] \) simply \( x = \tan \left( F - \frac{\pi}{2} \right) \)
  - efficiency of the acceptance is \( \frac{N_{\text{accepted}}}{N_{\text{MAX}}} = \int p(x)/ \int f(x) \)
    \( = 1/\pi \approx 0.32 \), i.e. efficiency can be increased by choosing
    \( f(x) = \frac{1}{2} \frac{1}{x^2 + 1} \), then \( x = \tan \left( 2F - \frac{\pi}{2} \right) \rightarrow 63\% \) acceptance
Alternative choice I: \( f(x) = \exp(-x) \) only for \( x \geq 0 \), then

- the integral \( F(x) \) is \( \int_0^x = -\exp(-x) + 1 \)
- the total area \( \int_0^\infty \exp(-x) \, dx = 1 > 0.5 = \int_0^\infty p(x) \)
- the inverse is \( x = -\log(-x + 1) \)
- to obtain also negative \( x \rightarrow \) add random sign \( \pm \)
Alternative choice II: \( f(x) = 1.1 \cdot \max(p(x)) \) in the compact interval \([0, 3]\), then

- it is \( \max(p(x)) = \frac{1}{\sqrt{2\pi}} \) in \([0, 3]\)
  \[\rightarrow f(x) = \frac{1.1}{\sqrt{2\pi}} \text{ in } [0, 3] \]
- hence \( F(x)^{-1} \) is \( x = \frac{F\sqrt{2\pi}}{1.1} - 0 \).
- the total area \( A \) is \( \frac{1.1}{\sqrt{2\pi}} \cdot (3 - 0) \)

\( \rightarrow \) clear: this choice (const. function) works only for compact intervals, otherwise \( A \) is infinite and \( F(x)^{-1} \) does not exist
Random walk
Random walk I

Idea: Brownian motion, e.g., dust in water (lab course: determination of diffusion coefficient \( D = \frac{\langle x^2 \rangle}{2t} \))

frequent collisions between dust particles and water molecules
→ frequent change of direction
→ trajectory not predictable even for few collisions
→ motion of dust particle into any direction with same probability

→ Random walk
like “drunken sailor”: \( N \) steps of equal length in arbitrary direction will lead to which distance from start point?
Random walk II

In one dimension:

- let’s start at \( x = 0 \), each step with length \( \ell \)
- for each step: probability \( p \) for step to the right and \( q = 1 - p \) to the left (independent from previous step)
- displacement after \( N \) steps

\[
x(N) = \sum_{i=1}^{N} s_i \quad \text{where } s_i = \pm \ell \quad \rightarrow \quad x^2(N) = \left( \sum_{i=1}^{N} s_i \right)^2
\]  

(35)

- for \( p = q = 1/2 \) \( \rightarrow \) coin flipping
- for large \( N \): \( \langle x(N) \rangle = 0 \) expected
but for $\langle x^2(N) \rangle$? $\rightarrow$ rewrite Eq. (35)

$$x^2(N) = \sum_{i=1}^{N} s_i^2 + \sum_{i \neq j=1}^{N} s_i s_j \quad (36)$$

where (for $i \neq j$) $s_i s_j = \pm \ell^2$ with same probability, so: $\sum_{i \neq j}^{N} s_i s_j = 0$

because of $s_i^2 = \ell^2 \rightarrow \sum_{i=1}^{N} s_i^2 = N\ell^2$:

$$\langle x^2(N) \rangle = \ell^2 N \quad (37)$$

evenly for constant time intervals of the random walk

$$\langle x^2(t) \rangle = \frac{\ell^2}{\Delta t} N \Delta t \left( = \frac{\ell^2}{\Delta t} t \right) \quad (38)$$

generally if $p \neq 1/2$ and $p$ for $+\ell$

$$\langle x(N) \rangle = (p - q)\ell N \quad (39)$$

$\rightarrow$ linear dependance on $N$
Example: Diffusion of photons in the Sun

Simplification: constant density, only Thomson scattering (free $e^-$) with cross section $\sigma_T = 6.652 \times 10^{-25} \text{ cm}^2$

mean free path length:

$$\ell = \frac{1}{n\sigma_T} = \left( \frac{\varrho}{m_H \sigma_T} \right)^{-1}$$ (40)

one dimension $\rightarrow$ only $R = R_\odot$, total time $t = N\Delta t$

$$\Rightarrow t = 9 \times 10^{10} \text{ s} = 2900 \text{ a} \ll t_{KH} (= 3 \times 10^7 \text{ a})$$
Importance of the random walk model

Many processes can be described by differential equation similar to diffusion equation (e.g., heat equation)

\[ \frac{\partial P(x, t)}{\partial t} = D \frac{\partial^2 P(x, t)}{\partial x^2} \]  

(41)

with diffusion coefficient $D$ and probability $P(x, t)dx$ to find particle at time $t$ in $[x, dx]$.

In 3 dimensions: $\frac{\partial^2}{\partial x^2} \equiv \nabla^2$

Moments: mean value of a function $f(x)$

\[ \langle f(x, t) \rangle = \int_{-\infty}^{+\infty} f(x, t)P(x, t)dx \]  

(42)

\[ \Rightarrow \langle x(t) \rangle = \int_{-\infty}^{+\infty} xP(x, t)dx \]  

(43)
Random walk VI

Compute integral in Eq. (43) → multiply Eq. (41) by $x$ and integrate over $x$

$$\int_{-\infty}^{+\infty} x \frac{\partial P(x, t)}{\partial t} dx = D \int_{-\infty}^{+\infty} x \frac{\partial^2 P(x, t)}{\partial x^2} dx$$

(44)

left hand side

$$\int_{-\infty}^{+\infty} x \frac{\partial P(x, t)}{\partial t} dx = \frac{\partial}{\partial t} \int_{-\infty}^{+\infty} x P(x, t) dx = \frac{\partial}{\partial t} \langle x \rangle$$

(45)

right hand side via integration by parts, note that $P(x = \pm \infty, t) = 0$, as well as all spatial derivatives:

$$D \int_{-\infty}^{+\infty} x \frac{\partial^2 P(x, t)}{\partial x^2} dx = Dx \left. \frac{\partial P(x, t)}{\partial x} \right|_{x=-\infty}^{x=+\infty} - D \int_{-\infty}^{+\infty} \frac{\partial P(x, t)}{\partial x} dx$$

(46)

$$= 0 - D P(x, t) \bigg|_{x=+\infty}^{x=-\infty} = 0$$

(47)

$$\Rightarrow \frac{\partial}{\partial t} \langle x \rangle = 0$$

(48)

I.e. $\langle x \rangle \equiv \text{const. for all } t$. For $x(t = 0) = 0 \rightarrow \langle x \rangle = 0$ for all $t$. 
Analogously for \( \langle x^2(t) \rangle \): integration by parts twice

\[
\frac{\partial}{\partial t} = 2D \int_{-\infty}^{+\infty} P(x, t) \, dx = 2D
\]

\[
\rightarrow \langle x^2(t) \rangle = 2D \, t
\]

compare with Eq. (38) \( \langle x^2(t) \rangle = \frac{\ell^2}{\Delta t} N \Delta t = \frac{\ell^2}{\Delta t} \, t \)

\( \rightarrow \) random walk and diffusion equation have same time dependence (linear)

(with \( 2D = \frac{\ell^2}{\Delta t} \))
Random numbers
for scientific purposes

- fast method to generate huge number of “random numbers”
- sequence should be reproducable

→ use deterministic algorithm to generate pseudorandom numbers

**Linear congruential method**

start with a seed \(x_0\), use one-dimensional map

\[
x_n = (ax_{n-1} + c) \mod m
\]  

(51)

- with integers \(a\) (multiplier), \(c\) (increment), \(m\) (modulus)
- \(m\) largest possible integer from Eq. (51) \(\rightarrow\) maximum possible period is \(m\) \(\rightarrow\) obtain \(r \in [0, 1)\) by \(x_n/m\)
- real period depends on \(a\), \(c\), \(m\), e.g., \(a = 3\), \(c = 4\), \(m = 32\), \(x_0 = 1\) \(\rightarrow\) 1, 7, 25, 15, 17, 23, 9, 31, 1, 7, 25, … \(\rightarrow\) period is 8 not 32
Better randomness can be obtained from physical processes:

- nuclear decay (*real* randomness!), e.g., → measure $\Delta t$ (difficult to implement)
- image noise, thermal noise (Johnson-Nyquist noise), e.g., → darkend USB camera (simple), special expansion cards with a diode
- “activity noise” in Unix:
  
  /dev/random
  /dev/urandom

  → random *bit* patterns from input/output streams (entropy pool) of the computer
  
  /dev/random blocks, if entropy pool is exhausted

For readout of Unix random devices need to interpret random bits(!) as numbers
Reading from urandom

E.g., by using fstream and union

```cpp
ifstream fin("/dev/urandom/");
union {unsigned int num;
    char buf[sizeof(unsigned int)];} u;
fin.read(u.buf, sizeof(u.buf));
cout << u.num;
```

→ fstream reads only char, buf and num are at the same address → read bits in as char output as unsigned int
quality check for uniformly distributed random numbers

- **equal distribution**: random numbers should be fair
- **entropy**: bits of information per byte of a sequence of random numbers (same as equal distribution)
- **serial tests**: for $n$-tuple repetitions (often only for $n = 2$, $n = 3$)
- **run test**: for monotonically increasing/decreasing sequences, also for length of stay for a distinct interval
- and more ...

Be careful!

There is no necessary or sufficient test for the randomness of a finite sequence of numbers.

→ can only check if it is “apparently” random
→ testing for “clumping” of numbers

Test for doublets

- define a square lattice $L \times L$ and fill each cell at random:
- array $n(x, y)$ with discrete coordinates
- choose random $1 \leq x_i, y_i \leq L$ where $x_i, y_i$ consecutive numbers of random number sequence
- fill cell $n(x_i, y_i)$ (e.g. set boolean to true)
- repeat procedure $t \cdot L^2$ times, $t$ is MC time step
- similar to nuclear decay, therefore expected: fraction of empty cells $\propto \exp(-t)$
Simple correlation test

- just plot $x_{i+1}$ over $x_i$ → look for suspicious patterns

correlation plot for linear congruential method with bad parameters

same plot but for C++ `rand()` function
testing for randomness (also: numbers or detections)

$\rightarrow \chi^2$ test

- let $y_i$ the number of events in bin $i$ and $E_i$ the expectation value
- e.g., $N = 10^4$ random numbers, $M = 100$ bins $\rightarrow E_i = 100$ (numbers/bin)
- the $\chi^2$ value (with $y_i$ measured number of random numbers in bin $i$):

$$\chi^2 = \sum_{i=1}^{M} \frac{(y_i - E_i)^2}{E_i}$$  \hspace{1cm} (52)

measures the conformity of the measured and the expected distribution

- the individual terms in Eq. (52) should be $\leq 1$, so for $M$ terms
  $\chi^2 \leq M \rightarrow \text{reduced } \chi^2$ by dividing by $M \rightarrow \text{“minimum” red. } \chi^2 = 1$
- e.g., 5 independent runs yield $\chi^2 \approx 92, 124, 85, 91, 99 \rightarrow$ as expected, in general: $\chi^2$ should be small (but $\chi^2 = 0$ is suspicious)
Confidence

- need a quantitative measure that shows normal distribution of the “error” \((y_i - E_i)\) → chi-squared distribution

\[
p(x, \nu) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} x^{(\nu-2)/2} e^{-x/2}
\]  

(53)

where \(\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt\) and \(\Gamma(z+1) = z!\)  

(54)

→ cumulated \(\chi^2\) distribution \(P(x, \nu)\):

\[
P(x, \nu) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} \int_0^x t^{(\nu-2)/2} e^{-t/2} dt
\]

(55)

with degrees of freedom \(\nu\), here: \(\nu = M - 1 = 99\), because of constraint \(\sum_{i=1}^M E_i = N\)
chi-square distribution

For $\nu > 30$ is $\sqrt{2x - \sqrt{2\nu - 1}}$ approximately normally distributed, for $\nu > 100$ is $x$ approximately normally distributed with $E = \nu$ and $\sigma = \sqrt{2}\nu$
function \( Q(x, \nu) = 1 - P(x, \nu) \)

\[ \rightarrow \text{probability that } \chi^2 > x \]

we want to check: How likely to get a \( \chi^2 \) of, e.g., 124 (our largest measured \( \chi^2 \))? \( \rightarrow \) solve \( Q(x, \nu) = q \) (probability \( \chi^2 > x \) for given \( x, \nu \)) for \( x \), or look it up in tables

for \( \nu = M - 1 = 99 \)

<table>
<thead>
<tr>
<th>( x )</th>
<th>138.9</th>
<th>134.6</th>
<th>123.2</th>
<th>110.6</th>
<th>98</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q )</td>
<td>0.005</td>
<td>0.01</td>
<td>0.05</td>
<td>0.2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

for our case: 1 out of 5 runs (20%) had \( y_2 = 124 \), but \( Q(x, \nu) \) implies for \( x = 123 \) only 5%, i.e., 1 out of 20 runs with \( \chi^2 \geq 123 \)

therefore: confidence level < 95%, rather 80% (because of \( q = 0.2 \) for \( x = 111 \))

try to increase confidence level: more runs \( \rightarrow \) if still only 1 out 20 with \( \chi^2 > 123 \) \( \rightarrow \) confidence level at 95%