Monte Carlo I

In this lecture, we will return to the idea of Monte Carlo sampling and by going beyond the direct sampling idea of Ulam (which we used in the statistical determination of π and also the percolation problem) and learn about Markov chain Monte Carlo and one of the most influential algorithms of all time — the Metropolis algorithm of 1953.

We will learn about these Monte Carlo techniques by going back to a problem set, which you have seen in the "Computersphysike" before, and that is integration methods and Monte Carlo integrators.

Standard integration methods

A Riemannian integral $f(x)$ over an interval $[a,b]$ can be evaluated by replacing it by a finite sum:

$$\int_a^b f(x)dx = \sum_{i=1}^{N} f(a+i\Delta x)\Delta x + O(\Delta x^2)$$

where $\Delta x = \frac{a-b}{N}$. The discretization error decreases as $1/N$ for this simple formula. Better approximations are the trapezoidal rule

$$\int_a^b f(x)dx = \Delta x\left[\frac{1}{2}f(a) + \sum_{i=1}^{N-1} f(a+i\Delta x) + \frac{1}{2}f(b)\right] + O(\Delta x^2)$$

$1/N^2$

or the Simpson rule (Kepler'sche Fassregel)

$$\int_a^b f(x)dx = \frac{\Delta x}{3}\left[f(a) + \sum_{i=1}^{N/2} 4f(a+(2i-1)\Delta x) + \sum_{i=1}^{N/2-1} 2f(a+2i\Delta x) + f(b)\right] + O(\Delta x^4)$$

which converges/scales like $\boxed{N^{-4}}$.

In higher dimensions the convergence is much slower though. With N points in d dimensions the linear distance between two points scales only as $N^{-1/d}$.
Thus the Simpson rule in d dimensions converges only as $N^{-4/d}$, which is very slow for large d.

The solution are Monte Carlo integrators.

Phase space for classical N-body problem has dimension $6N$, as there are three coordinates each for the location and momentum of each particle.

Monte Carlo integrators

With randomly chosen points the convergence does not depend on dimensionality. Using N randomly chosen points X_i the integral can be approximated by

$$\frac{1}{\Omega}\int f(x)dx \approx \overline{f} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

where $\Omega = \int dx$ is the integration volume.

As we saw previously, the error of such a direct Monte Carlo sampling estimate scales as $N^{-1/2}$.

Thus, in dimensions $d \geq 9$ Monte Carlo methods are preferable to a Simpson rule.

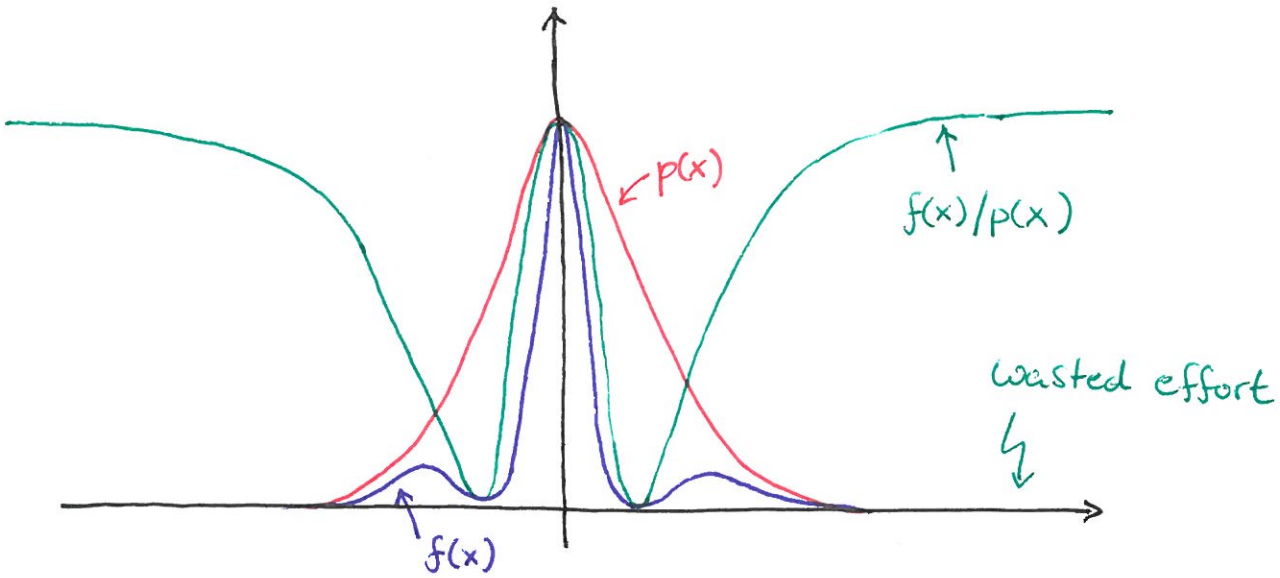
This simple Monte Carlo integration is however not the ideal method. The reason is the variance of the function

$$\mathrm{Var}\, f = \frac{1}{\Omega}\int f(x)^2 dx - \left[\frac{1}{\Omega}\int f(x)dx\right]^2 \approx \frac{N}{N-1}\left(\overline{f^2} - \overline{f}^2\right)$$

The error of Monte Carlo sampling (i.e. direct sampling) is

$$\Delta = \sqrt{\frac{\mathrm{Var}\, f}{N}} \approx \sqrt{\frac{\overline{f^2} - \overline{f}^2}{N-1}}$$

The problem of direct sampling is that oftentimes the function of interest is strongly peaked in a small region of phase space and has a large variance



Lots of "time" is wasted in regions where the function is small. The sampling error is large, since the variance is large.

The solution to this problem is "importance sampling", where the points x_i are chosen not uniformly but according to a probability distribution $p(x)$ with

$$\int p(x)dx = 1$$

Using these p-distributed random points the sampling is done according to

$$\langle f \rangle = \frac{1}{\Omega}\int f(x)dx = \frac{1}{\Omega}\int \frac{f(x)}{p(x)}\cdot p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N}\frac{f(x_i)}{p(x_i)}$$

for x_i chosen acc. to $p(x)$

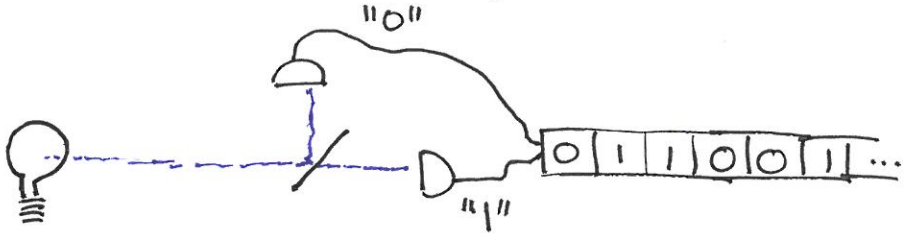
where the error now becomes

$$\Delta = \sqrt{\frac{\text{Var } f/p}{N}}$$

Thus, we want to find a distribution function p as similar to f as possible and such that p-distributed random numbers are easily available

Generating random numbers

Real random numbers are hard to obtain. Options to consider might include (classical) chaotic systems, such as atmospheric noise, or quantum mechanical systems such as the one used in a commercial product (idquantique.com):



The commercial USB device produces random numbers at a speed of 4 Mbit/s, which is too slow for most MC simulations.

Pseudo random numbers

Pseudo random numbers are generated by an algorithm, and as such they are not random at all, but completely deterministic. However, they "look" nearly random when the generating algorithm is not known.

Pseudo random numbers may be good enough for our purposes – however never trust pseudo random numbers!

Pseudo random number generators include linear congruential generators, which are of the simple form

$$X_{n+1} = f(X_n)$$

A reasonably good choice is the GGL generator

$$X_{n+1} = (a X_n + c) \bmod m$$

with $a = 16\,807$ ← 7^5, $c = 0$, $m = 2^{31} - 1$ ← Apple Carbon Lib. The quality depends sensitively on this choice of (a, c, m).

However, for such a 32-bit generator periodicity is a problem – the sequence repeats identically after $2^{31}-1$ iterations. With 500 million numbers per second that is just 4 seconds. Thus, these methods should not be used anymore.

A better alternative are lagged Fibonacci generators of the form

$$X_n = X_{n-p} \otimes X_{n-q} \bmod m$$

with $p > q$ and an initial set of elements $X_1, X_2, \ldots, X_p$.
$\otimes$ is one of the following binary operations $+, -, \times, \oplus$ (exclusive or)
m is usually a power of 2, often 2^{32} or 2^{64}, let's define 2^M.

The maximum period is

$(2^p - 1) \cdot 2^{M-1}$ for $+, -$

$(2^p - 1) \cdot p$ for $\oplus$

$(2^p - 1) \cdot 2^{M-3}$ for $\times$ (or 1/4 the period of add. case)

Good choices for $(p, q, \otimes)$ are

$(2281, 1252, +)$ $(9689, 5502, +)$ $(44497, 23463, +)$

All of these have extremely long periods due to the large block of seeds. These seed blocks are typically generated by linear congruential (see above). The lagged Fibonacci generator is also a very fast generator: it vectorizes and pipelines very well.

More advanced generators rely more heavily on number theory, including the Mersenne twister (Matsumoto & Nishimura, 1997) and the WELL generator (Panneton & L'Ecuyer, 2004).

Non-uniform random numbers

We have now seen a number of ways to generate pseudo random numbers u in the interval $[0,1[$.

From this we can easily obtain uniform distributions in other intervals, e.g. uniform x in $[a,b[$: $x = a + (b-a)\cdot u$.

For other distributions we can rely on two approaches:

- inversion of integrated distribution
- rejection method

Inversion method

How can we get a random number x distributed with $f(x)$ in the interval $[a,b[$ from a uniform random number u?

- Look at probabilities

$$P[x<y] = \int_a^y f(t)\,dt \Big/ \int_a^b f(t)\,dt =: F(y) \equiv P[u<F(y)]$$

$$\Rightarrow x = F^{-1}(u)$$

This method is feasible if the integral can be inverted easily, e.g. an exponential distribution $f(x) = \lambda \exp(-\lambda x)$ can be obtained from a uniform one by

$$x = -\frac{1}{\lambda} \ln(1-u).$$

The normal distribution $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2)$ cannot easily be integrated in one dimension, but can be easily integrated in two dimensions.

We can, however, obtain two normally distributed numbers from two uniform ones (Box-Muller method)

$$u_1 = \sqrt{-2\ln(1-u_1)}\,\sin u_2 \qquad u_2 = \sqrt{-2\ln(1-u_1)}\,\cos u_2$$

Rejection method (von Neumann)

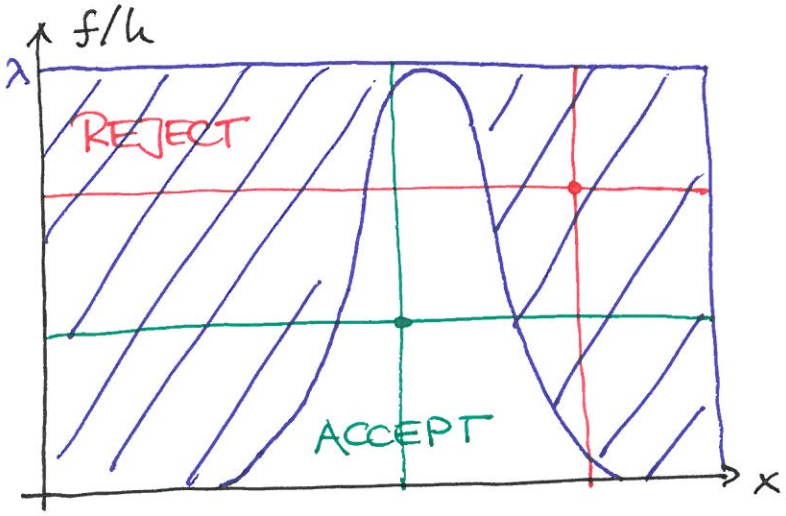
Look for a simple distribution h that bounds f

$$f(x) < \lambda \cdot h(x)$$

(λ: a constant (needed since both $f(x)$ and $h(x)$ are normalized))

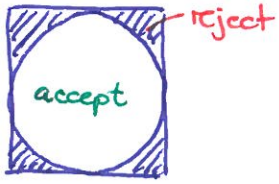
- Choose an h-distributed number x
- Choose a uniform random number $0 \leq u < 1$
- Accept x if $u < \dfrac{f(x)}{\lambda h(x)}$

otherwise reject x and get a new pair of random numbers.



(rectangular shape for uniform proposal distribution $h(x)$)

Note that it is precisely this rejection method which we used to statistically sample the value of π:



Markov chain Monte Carlo

In contrast to the percolation problem or the simple exponential distribution(s) discussed before it will in general not be possible to directly create p-distributed random points. Instead a Markov process can be used.

Starting from an initial point x_0 a Markov chain of states is generated

$$X_0 \to X_1 \to X_2 \to \ldots \to X_n \to X_{n+1} \to \ldots$$

no memory!

A transition matrix $\boxed{W_{xy}}$ gives the transition probabilities of going from state x to state y in one step of the Markov process. As the sum of probabilities of going from state x to any other state is one, the colums of the matrix W are normalized

$$\sum_y W_{xy} = 1$$

A consequence is that the Markov process conserves the total probability. Another consequence is that the largest eigenvalue of the transition matrix W is 1 and the corresponding eigenvector with only positive entries is the equilibrium distribution which is reached after a large number of Markov steps.

We want to determine the transition matrix W so that we asymptotically reach the desired probability $p(x)$ for a configuration x_i.
A set of sufficient conditions is:

- <u>normalization</u>:

$$\sum_y W_{x,y} = 1$$

- <u>ergodicity</u>: It has to be possible to reach any configuration x from any other configuration y in a finite number of Markov steps. This means that for all x and y there exists a positive integer $n<\infty$ such that $(W^n)_{xy} \neq 0$

$$\forall x,y \; \exists n : (W^n)_{xy} \neq 0$$

- <u>balance</u>: The probability distribution $P_x^{(n)}$ changes at each step of the Markov process:

$$\sum_x P_x^{(n)} W_{x,y} = P_y^{(n+1)} ,$$

but converges to the equilibrium distribution P_x.
The equilibrium distribution P_x is an eigenvector of the transition matrix with left eigenvalue 1 and the <u>equilibrium condition</u>

$$\sum_x P_x W_{x,y} = P_y$$

must be fulfilled.
It is easy to see that the <u>detailed balance condition</u>

$$\frac{W_{x,y}}{W_{y,x}} = \frac{P_y}{P_x} \quad \Leftrightarrow \quad P_x W_{x,y} = P_y W_{y,x}$$

is sufficient.

One way to fullfill the detailed balance condition is to use a so-called heat-bath approach (Gibbs weights)

$$W_{x,y} = \frac{P_y}{P_x + P_y} \qquad W_{y,x} = \frac{P_x}{P_x + P_y}$$

$$\Rightarrow \frac{W_{x,y}}{W_{y,x}} = \frac{P_y}{P_x} \quad \checkmark$$

The Metropolis algorithm

The simplest incarnation for a Markov chain Monte Carlo algorithm comes in the form of the Metropolis algorithm, which was developed by Metropolis, Rosenbluth[2] and Teller[2] at Los Alamos National Laboratory in 1953.

Here is the idea:

- Starting with a point x_i choose randomly one of a fixed number N of changes Δx, and propose a new point $x' = x_i + \Delta x$.
- Calculate the ratio of the probabilities $P = \frac{P_{x'}}{P_{x_i}}$
- If $P > 1$, the next point is $x_{i+1} = x_i + \Delta x = x'$
- If $P < 1$, the next point is $x_{i+1} = x'$ with probability P, otherwise $x_{i+1} = x_i$ (rejection method).

  (We do this by drawing a random number u uniformly distributed in $[0,1[$ and set $x_{i+1} = x'$ if $u < P$, else $x_{i+1} = x_i$.)
- Measure a quantity of interest / add to your integration. Go back to step 1.

The algorithm is ergodic if one ensures that the N possible random changes allow all points in the integration domain to be reached in a finite number of steps.

If in addition for each change Δx there is also an inverse change $-\Delta x$, we fulfill detailed balance

$$\frac{W_{ij}}{W_{ji}} = \frac{\frac{1}{N}\min\left(1, \frac{P(j)}{P(i)}\right)}{\frac{1}{N}\min\left(1, \frac{P(i)}{P(j)}\right)} = \frac{P(j)}{P(i)} \quad .$$

Autocorrelation effects

In contrast to a direct sampling method two successive points x_i, x_{i+1} in a Markov chain are not fully independent but correlated. These correlations between configurations also manifest themselves in correlations between the measurements of a quantity A measured in the Monte Carlo process. We will need to take these autocorrelations into account when we determine the statistical errors of our Monte Carlo estimates.

Denote by $\boxed{A(t)}$ the measurement of the observable A evaluated at the t-th Monte Carlo point x_t.
The autocorrelations decay exponentially for large time differences

$$\langle A_t A_{t+\Delta} \rangle - \langle A \rangle^2 \propto \exp\left(-\frac{\Delta}{\boxed{\tau_A^{(exp)}}}\right)$$

Note that the autocorrelation time τ_A depends on the quantity A

An alternative definition is the integrated autocorrelation time defined by

$$\tau_A^{(int)} = \frac{\sum_{\Delta=1}^{\infty} \left(\langle A_t A_{t+\Delta} \rangle - \langle A \rangle^2 \right)}{\langle A^2 \rangle - \langle A \rangle^2}$$

As usual the expectation value of the quantity A can be estimated by the mean $\overline{A}$ (analogous to the case of truly independent measurements in the direct sampling approach).

The error estimate, however, needs to be modified. It is given by the expectation value of the squared difference between sample average and expectation value:

$$(\Delta A)^2 = \left\langle \left(\overline{A} - \langle A \rangle \right)^2 \right\rangle$$

$$= \left\langle \left(\frac{1}{N} \sum_{t=1}^{N} A(t) - \langle A \rangle \right)^2 \right\rangle$$

$$= \left\langle \frac{1}{N^2} \sum_{i=1}^{N} \left(A(t)^2 - \langle A \rangle^2 \right) \right\rangle$$

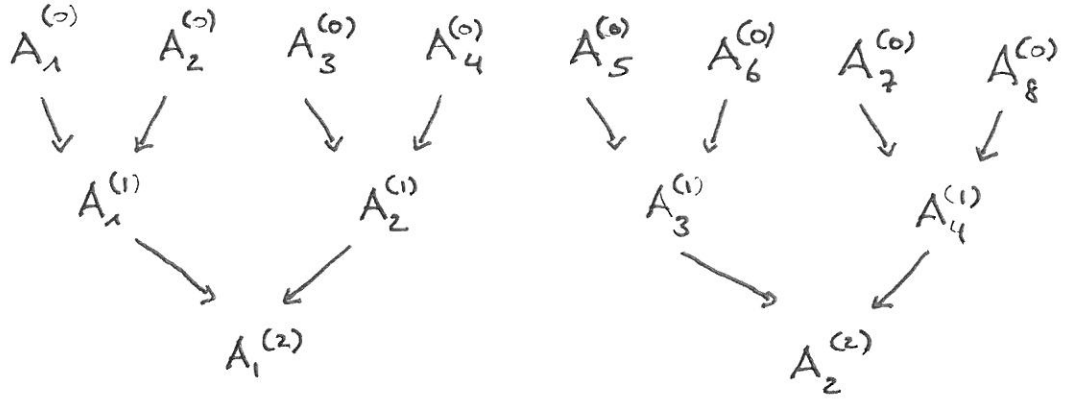
$$+ \frac{2}{N^2} \sum_{t=1}^{N} \sum_{\Delta=1}^{N-t} \left(\left\langle A(t) A(t+\Delta) \right\rangle - \left\langle \langle A \rangle^2 \right\rangle \right)$$

$$\approx \frac{1}{N} \operatorname{Var} A \left(1 + 2 \tau_A^{(int)} \right)$$

In particular, we see that the number of statistically uncorrelated samples is reduced from N to $\frac{N}{1 + 2\tau_A^{(int)}}$.

The binning analysis

The binning analysis is a reliable way to estimate the integrated autocorrelation times. Starting from the original series of measurements $A_i^{(0)}$ with $i = 1, \dots, N$ we iteratively create "binned" series by averaging over consecutive entries



or in equations

$$A_i^{(e)} := \frac{1}{2}\left(A_{2i-1}^{(e-1)} + A_{2i}^{(e-1)} \right) \qquad \text{with } i = 1, \dots, N_e = \frac{N}{2^e}$$

These <u>bin averages</u> $A_i^{(e)}$ are <u>less correlated</u> than the original values $A_i^{(0)}$. The <u>mean value</u> remains the <u>same</u>.

The <u>errors</u> $\Delta A^{(e)}$ estimated incorrectly from the variance

$$\Delta A^{(e)} = \sqrt{\frac{\mathrm{Var}\, A^{(e)}}{N_e - 1}} \approx \frac{1}{N_e} \sqrt{\sum_{i=1}^{N_e} \left(A_i^{(e)} - \overline{A^{(e)}} \right)^2}$$

however <u>increase</u> as a function of bin size 2^e.

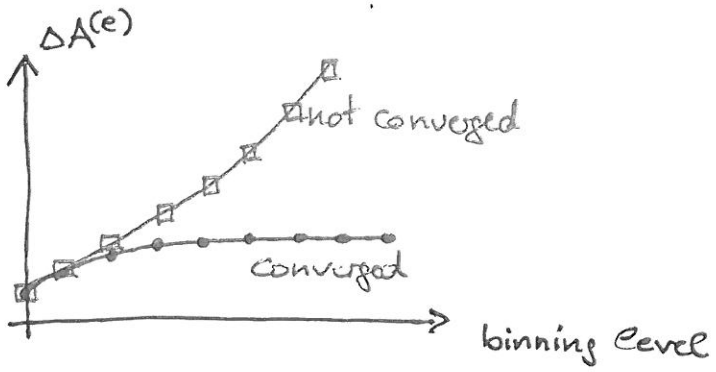
For $\boxed{2^e \gg \tau_A^{(\mathrm{int})}}$ the <u>bins</u> become <u>uncorrelated</u> and the <u>errors</u> <u>converge</u> to the correct estimate

$$\Delta A = \lim_{e \to \infty} \Delta A^{(e)}$$

that is

$$\Delta A^{(e)} = \sqrt{\frac{\mathrm{Var}\, A^{(e)}}{N_e - 1}} \xrightarrow{e \to \infty} \Delta A = \sqrt{\frac{\mathrm{Var}\, A}{N-1} \left(1 + 2\tau_A^{(\mathrm{int})} \right)}$$

The binning analysis thus gives a reliable recipe for estimating errors and autocorrelation times.
One has to calculate the error estimates for different bin sizes ℓ and check if they converge to a limiting value.



If convergence is observed the limit ΔA allows to obtain an estimate for the autocorrelation time $\tau_A^{(int)}$

$$\tau_A^{(int)} = \frac{1}{2}\left[\left(\frac{\Delta A}{\Delta A^{(0)}} \right)^2 - 1 \right]$$

Equilibration

Equilibration / thermalization is as important as autocorrelations. The Markov chain converges only asymptotically to the desired distribution. Consequently, Monte Carlo measurements should be started only after a large number N_{eq} of equilibration steps, when the sampled distribution is sufficiently close to the asymptotic distribution.
N_{eq} has to be much larger than the thermalization time $\tau_A^{(eq)}$ which is defined similar to the autocorrelation time as

$$\tau_A^{(eq)} = \frac{\sum_{\Delta=1}^{\infty} \left(\langle A_0 A_\Delta \rangle - \langle A \rangle^2 \right)}{\langle A_0 \rangle \langle A \rangle - \langle A \rangle^2}$$

It can be shown that the thermalization time is the maximum of all autocorrelation times for all observables.

The equilibration time is in fact related to the second largest eigenvalue of the transition matrix

$$\tau^{(eq)} = -\frac{1}{\ln \Lambda_2} .$$

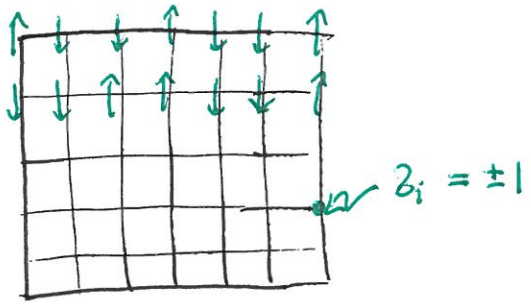
Λ_2 ← 2nd largest eigenvalue of transition matrix

It is recommended to thermalize the system at least a hundred times the thermalization time before starting measurements.

The Ising model

The Ising model is the simplest model for a magnetic system and a prototype statistical system. We will use it for our discussion of thermodynamic phase transitions. In doing so, we will face the dynamic problem of evaluating thermodynamic averages through phase space integrals — thus moving beyond the 'static' problem of percolation.

The Ising model consists of an array of classical spins $S_i = \pm 1$ that can point either up ($S_i = +1$) or down ($S_i = -1$).



The Hamiltonian is

$$H = -J \sum_{\langle i,j \rangle} S_i S_j \,.$$

$\langle i,j \rangle$: Sum over nearest neighbors

Two parallel spins contribute an energy $-J$, while two antiparallel spins contibute $+J$. In the ferromagnetic case ($J>0$) the state of lowest energy is the fully polarized state where all spins are aligned, either pointing up or down. ($T=0$)

At finite temperatures the spins start to fluctuate and also states of higher energy contribute to thermal averages.
The average magnetization thus decreases from its full value at zero temperature.

At a critical temperature T_c there is a second order phase transition to a disordered phase — similar to the phase transition we discussed in the percolation problem.

The Ising model is the simplest magnetic model exhibiting such a phase transition and is often used as a prototype model for magnetism. To discuss the phase transition we will again use the scaling hypothesis introduced for the percolation transition.

From statistical mechanics (and the corresponding lecture) we know how to calculate the thermal average of a quantity A at a finite temperature T by summing over all states

$$A(T) = \frac{1}{Z} \sum_i A_i \exp(-\beta E_i)$$

A_i, E_i: values of A and E for configuration i

$\beta = \frac{1}{k_B T}$ is the inverse temperature

where Z is the partition function ("Zustandssumme")

$$Z = \sum_i \exp(-\beta E_i)$$

which normalizes the probabilities $p_i = \frac{\exp(-\beta E_i)}{Z}$

While it is possible to evaluate these sums exactly for small systems, we will turn to Monte Carlo summation/integration for larger systems for which the number of states grows exponentially like 2^N.

The single spin flip Metropolis algorithm

As we have seen before it will not be efficient to use direct sampling to estimate thermal averages via Monte Carlo. Instead we will again turn to importance sampling – where the states i are not chosen uniformly but with the correct probability $p_i = \exp(-\beta E_i)/Z$ – which we implement via a Markov chain Metropolis algorithm.

We construct a Markov chain in phase space by

- starting with a configuration c_i propose to flip a single spin, leading to a new configuration c'
- calculate the energy difference $\Delta E = E(c') - E(c_i)$ between configurations c' and c_i.
- if $\Delta E < 0$ the next configuration is $c_{i+1} = c'$.
- if $\Delta E > 0$ then $c_{i+1} = c'$ with probability $\exp(-\beta \Delta E)$, otherwise $c_{i+1} = c_i$.

  We do this by drawing a random number u uniformly distributed in the interval $[0, 1[$ and set $c_{i+1} = c'$ if $u < \exp(-\beta \Delta E)$.
- measure all the quantities of interest in the new configuration.

The algorithm is ergodic, since any configuration can be reached from any other one in a finite number of spin flips.

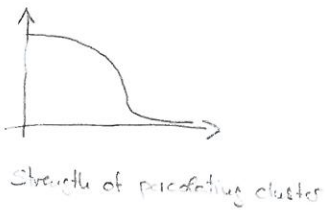
It also fulfills the detailed balance condition.

Critical behavior of the Ising model

Close to the thermal phase transition at $\boxed{T_c = \frac{2}{\ln(1+\sqrt{2})} \approx 2{,}27}$ again scaling laws charackrize the behavior of all physical quantities.

The average magnetization scales as

$$m(T) = \left\langle \frac{|M|}{V} \right\rangle \propto (T_c - T)^{\beta}$$



where $M = \sum_i \sigma_i$ is the total magnetization and V the system volume (= number of spins).

The magnetic susceptibility $\chi = \left.\frac{dm}{dh}\right|_{h=0}$ can be calculated from magnetization fluctuations and diverges with the exponent γ

$$\chi(T) = \frac{\langle M^2/V^2 \rangle - \langle |M|/V \rangle^2}{T} \propto |T_c - T|^{-\gamma}$$

The correlation length ξ is defined by the asymptotically exponential decay of the two-spin correlations

$$\langle \sigma_0 \sigma_r \rangle - \langle |m| \rangle^2 \propto \exp(-r/\xi)$$

magnetization per site

It is best calculated from the structure factor $S(\vec{q})$, defined as the Fourier transform of the correlation function.
For small $\vec{q}$ the structure factor has a Lorentzian shape

$$S(\vec{q}) = \frac{1}{1+q^2\xi^2} + O(q^4)$$

The correlation length diverges as

$$\xi(T) \propto |T - T_c|^{-\nu}$$

At the critical temperature the correlation function again follows the same power law as in the percolation problem

$$\langle s_0 s_r \rangle \propto r^{-(d-2+\eta)}$$

where $\eta = \frac{2\beta}{\nu} - d + 2$ can be derived from scaling laws as in the percolation problem.

The specific heat $C_V(T) = \frac{dE}{dT}$ can be calculated from the energy fluctuations

$$C_V(T) = \frac{\langle E^2/V^2 \rangle - \langle E/V \rangle^2}{T^2} \propto |T - T_c|^{-\alpha} \propto \ln |T - T_c|$$

H diverges logarithmically in two dimensions, since $\alpha = 0$.

cluster density

Like in percolation, finite-size scaling is the method of choice for the determination of these exponents.

A good estimate for T_c is obtained from the Binder cumulant

$$U = 1 - \frac{\langle M^4 \rangle}{3\langle M^2 \rangle^2}$$

which has a universal value (independent of system size) at T_c – just like the probability $\Pi(p, L)$ to find a percolating cluster in the percolation problem.

This universal value of the Binder cumulant is independent of system size and the curves $U(T)$ for different system sizes L all cross in one point at T_c.

This is a consequence of the finite-size scaling ansatz

$$\langle M^4 \rangle = (T-T_c)^{4\beta}\, u_4\left((T-T_c)L^{1/\nu}\right)$$

$$\langle M^2 \rangle = (T-T_c)^{2\beta}\, u_2\left((T-T_c)L^{1/\nu}\right)$$

which gives

$$U(T,L) = 1 - \frac{u_4\left((T-T_c)L^{1/\nu}\right)}{3\,u_2\left((T-T_c)L^{1/\nu}\right)^2}$$

which for $T=T_c$ is universal and independent of system size

$$U(T_c,L) = 1 - \frac{u_4(0)}{3u_2(0)^2}$$

High-precision Monte Carlo simulations actually show that not all lines cross exactly at the same point, but that due to higher order corrections to the finite-size scaling ansatz the crossing point moves slightly, proportional to $L^{-1/\omega}$, allowing a high precision estimate of T_c and ν.