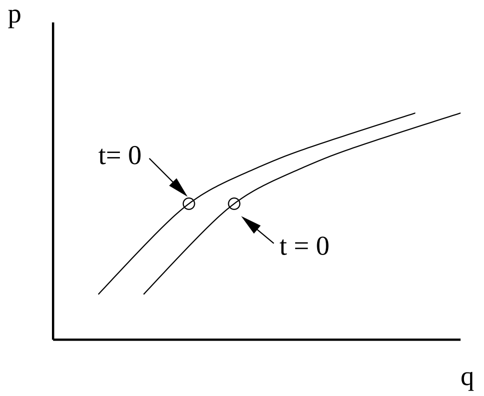

Figure 1: Two trajectories in phase-space

Abstract

Statistical Physics (part 2), the original Metropolis Algorithm, Simulated Annealing.

Phase Space

It is convenient to visualize a mechanical system as a point in the $6N$ dimensional space (q, p) of all the positions and momenta of all the N particles.

Due to the complexity of macroscopic systems ($N \sim 10^{24}$) it was necessary to abandon determinism and use statistics to describe the system. The predictions of statistical physics are expected to hold only on the average.

Instead of the precise initial conditions (which are unknown), statistical physics describes the system by a probability distribution over phase space, $\rho(q, p)$ for $t = 0$. As it will be seen later, Hamilton's equations imply the conservation, at all time, of this initial distribution. This is the famous, Liouville's theorem. The determination of ρ is then the first step.

Maxwell-Boltzmann-Gibbs Distribution

Different forms for ρ are found to be needed depending on the particular data available about the system. We will be concerned only with the so called canonical distribution. We assume that the system is not isolated but in thermal equilibrium with a heat bath at constant temperature T. Statistically this is equivalent to the assumption that the average energy of the molecules is constant. The novel idea of Boltzmann was to discretize phase-space to find the most likely distribution for ρ.

Each particle has a definite position and momentum. Subdivide the positions and momentums for each particle into m (6 dimensional) cells of equal size.

Assume that these cells are small enough so that the value of the energy within each cell is approximately constant. Let E_j be the energy in the j-th cell. Assume further that the cells, even though small, they are still big enough to accommodate lots of particles inside. These are reasonable assumptions justified by the smallness of atomic dimensions ($\sim 10^{-8}$ cm), the size of typical N and the smoothness of energy surfaces. This discretization of the phase-space for each molecule into m equal size cells induces a discretization of the phase-space of the system into, m^N equal size cells. With the help of this discretization, the state s of the system is specified by,

$$s \in \{1, 2, \dots, m\}^N$$

indicating the cell number for each of the N particles. If the particles are assumed to be identical and indistinguishable, then permutations of the molecules with a given cell number have no physical consequences. All it matters is how many molecules end up in each of the cells and not which ones did. Thus, the actual set of distinguishable physical states is much smaller than m^N it is,

$$\frac{(N + m - 1)!}{N!(m - 1)!}$$

corresponding to the number of ways of splitting N into the m cells. There are,

$$G = \frac{N!}{n_1! n_2! \dots n_m!}$$

ways of throwing the N molecules into the m cells in such a way that n_1 of them are in the first cell, n_2 in the second cell, etc. If we assume a priori that the molecules have equal chance of ending in any of the cells then the number of ways can be turned into a probability for the state $s = (n_1, \dots, n_m)$,

$$P = \frac{N!}{n_1! n_2! \dots n_m!} \times \text{constant}$$

Hence, the most likely distribution of balls among the cells is the one that maximizes this probability subject to whatever is known about the system. When the temperature is all we know we maximize P subject to the constraint that the average energy is fixed at kT. Where k is a phenomenological (not fundamental) constant needed to change the units from *ergs* (units of energy) to *degrees* (usual units for temperature). It is known as the Boltzmann constant and it is about,

$$k = 1.380 \times 10^{-14} \text{ergs per degree centigrade}$$

Using the fact that N and the n_j are large we can use Stirling's approximation,

$$\log n! \sim n \log n - n$$

to get,

$$
\begin{aligned}
\log P &= N \log N - \sum_j (n_j \log n_j - n_j) + \text{constant} \\
&= -N \sum_j p_j \log\ p_j + \text{constant}
\end{aligned}
$$

where,

$$p_j = \frac{n_j}{N}.$$

Thus, P is the probability of observing the probability distribution $(p_1, \ldots, p_m)$. A probability of a probability... A prior!

$$P \propto e^{-N \sum_j p_j \log\ p_j}$$

Known as an entropic prior, for the quantity in the exponent (sans N) is the famous expression for the entropy of a probability distribution. If we treat the p_j as if they were continuous variables we can obtain the most likely a priori distribution by solving,

$$
\begin{aligned}
&\max_{s.t.} - \sum_j p_j \log\ p_j \\
&\sum_j p_j = 1 \\
&\sum_j p_j E_j = kT
\end{aligned}
$$

Using Lagrange multipliers for the constraints we can find the maximum by maximizing,

$$\mathcal{L} = \sum_j p_j \log\ p_j - \alpha - \sum_j p_j - \beta \sum_j p_j E_j$$

Taking derivatives,

$$\frac{\partial \mathcal{L}}{\partial p_i} = 0 \Rightarrow \log\ p_i - 1 - \beta E_i - \alpha = 0$$

from where we get,

$$p_i = \frac{1}{Z} e^{-\beta E_i}$$

where the normalization constant,

$$Z = \sum_i e^{-\beta E_i}$$

is known as the partition function. In order to satisfy the constraint of average energy we need to take,

$$\frac{\sum_i E_i e^{-\beta E_i}}{\sum_i e^{-\beta E_i}} \approx \int_0^\infty E\beta e^{-\beta E} dE = kT$$

and since the middle integral is the mean of the exponential distribution we get,

$$\beta = \frac{1}{kT}.$$

The Original Metropolis Algorithm: Circa 1953

It was proposed as an algorithm to simulate the evolution of a system in a heat bath towards thermal equilibrium. From a given state i of energy E_i, generate a new state j of energy E_j by a small perturbation, e.g. changing one of the position coordinates of one of the particles a little. If the new proposed state j has smaller energy than the initial state i then make j the new current state, otherwise accept state j with probability,

$$A_{ij}(T) = exp(-(E_j - E_i)/kT)$$

where T is the temperature of the heat bath. After a (possibly) large number of iterations we would expect the algorithm to visit states of different energies according to the canonical distribution. In fact this can be rigorously justified by showing that the sequence of states visited by the algorithm forms an ergodic Markov Chain with the canonical distribution as the stationary distribution for the chain.

Let, us get closer to the theory of Markov Chains by using the usual notation. Define,

$$X_t = \text{state of the system at time } t$$

The one step transition probabilities for the Metropolis (like) algorithm are,

$$p_{ij}(T) = P[X_{t+1} = j | X_t = i] = \begin{cases} G_{ij}(T)A_{ij}(T) & \text{if } i \neq j \\ 1 - \sum_{k \neq j} p_{ij}(T) & \text{if } i = j \end{cases}$$

where,

$$\begin{aligned} G_{ij}(T) &= \text{prob. of generating j from i} \\ A_{ij}(T) &= \text{prob. of accepting j from i} \end{aligned}$$

The acceptance Metropolis probabilities can be written as,

$$A_{ij}(T) = exp(-(E_j - E_i)_+/kT)$$

where,

$$x_+ = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Any, probability distribution π over the set of states satisfying the reversibility condition known as *detailed balance*,

$$p_{ij}\pi_i = p_{ji}\pi_j$$

will be a stationary distribution for the Markov Chain with transition probabilities p_{ij}. This can be easily seen by adding over j both sides of the previous equation,

$$\begin{aligned} P[X_t = j] &= \pi_i = \sum_j p_{ij}\pi_i \\ &= \sum_j p_{ji}\pi_i = P[X_{t+1} = j] \end{aligned}$$

It can be readily checked that when the generating probabilities are symmetric in i and j i.e. when,

$$G_{ij}(T) = G_{ji}(T)$$

we have detailed balance with,

$$\pi_i = \frac{1}{Z}\exp(-E_i/kT)$$

i.e. the canonical distribution. Just consider each of the cases separately. It is obviously true when $i = j$ and for $i \neq j$ with $E_i < E_j$ the detailed balance condition reduces to the simple identity,

$$\exp(-(E_j - E_i)/kT)\exp(-E_i/kT) = \exp(-E_j/kT)$$

the other case interchanges i with j, which is also obviously true. This does not show, however that the chain is ergodic, i.e. that the distribution of X_t will converge to the stationary canonical distribution.

Simulated Annealing

Annealing is a physical process often used in practice to get rid of cracks and impurities from a solid in order to increase its strength. This is done by first heating the solid until it melts and then slowly decreasing the temperature to allow the particles to re-arrange themselves in the state of lowest possible energy (ground state). The opposite of annealing is known as quenching. The solid is melted but then the temperature is quickly lowered so that the particles get frozen in a local minimum for the energy (meta-stable state). It is convenient to think of annealing as a way of using nature to solve a minimization problem in billions of variables. The annealing process is simulated by the Metropolis algorithm when we take a sequence of slowly decreasing temperatures converging to 0. If we run the Metropolis algorithm with each value of the temperature for a long time until it reaches the asymptotic cannonical distribution, then in the limit when the temperature approaches 0 the system will be found on state i with probability,

$$\begin{aligned} \lim_{T \to 0} \pi_i(T) &= \lim_{T \to 0} \frac{\exp\{-E_i/kT\}}{\sum_j \exp\{-E_j/kT\}} \\ &= \lim_{T \to 0} \frac{\exp\left\{\frac{E^*-E_i}{kT}\right\}}{\sum_j \exp\left\{\frac{E^*-E_i}{kT}\right\}} \end{aligned}$$

where,

$$E^* = \min_i E_i = \text{ Global min of energy}$$

Thus, the exponents (in the above ratios) are always either zero or negative. In the limit when $T \to 0$ the terms with negative exponents disappear and we get,

$$\lim_{T \to 0} \pi_i(T) = \begin{cases} \frac{1}{N^*} & \text{if } E_i = E^* \\ 0 & \text{otherwise} \end{cases}$$

where,

$$N^* = |\{i : E_i = E^*\}|$$

Thus, $\lim_{T \to 0} \pi_i(T)$ is uniformly distributed over the set of states of global minimum energy!

Simulated annealing is one of the few known algorithms assuring convergence to a global minimum. It is often used in combination with efficient steepest descent methods, such as conjugate gradients, as a way for avoiding getting trap in local minima. This is what the theory says but in practice the performance of annealing depends primarily on the cooling schedule, i.e. how exactly is the temperature decreased and second on the stopping criterion, i.e. how it is decided to stop the algorithm, for example how close to zero is the temperature allowed to go before stopping.