A Short Introduction to Statistical Physics

Carlo Lucibello and Enrico Malatesta

v1.2 — 2026-02-16

Abstract. Lecture notes for the first 6 lessons of the "Complex Systems and Physics Models" course at Bocconi University.

Contents

1 Logistic Details for the Course

Course materials and communications will be on Piazza.

The **Syllabus** is on Piazza as well.

Office hours: whenever you want, write an email.

Instructors: Carlo Lucibello (carlo.lucibello@unibocconi.it) and Enrico Malatesta (enrico.malatesta@unibocconi.it).

Teaching Assistant: Alessandra Passalacqua (alessandra.passalacqua@phd.unibocconi.it).

Bibliographic References:
- These notes.
- Mézard and Montanari, "Information, Physics, and Computation". Chapter 2.
- Krzakala and Zdeborová, "Statistical Physics methods in Optimization & Machine Learning: An Introduction to Replica, Cavity, & Message Passing Techniques". Chapter 1.

2 What is Statistical Physics?

Statistical physics started with the study of gases, trying to find a bridge between the microscopic dynamics of atoms and the macroscopic properties of gases investigated by thermodynamics (e.g. pressure and temperature).

Now it generically deals with systems with many entities, may that be electrons in superconductors or agents in markets, weights in neural networks, etc... .

A key concept is that of *equilibrium*: the equations describing systems at equilibrium do not have a time dependence and details of the microscopic dynamics are not important. Descriptions are given in terms of probability distributions.

In this course, we will not be interested in statistical physics as a tool to study "natural" systems, but we are rather interested in the methods and concepts that are used in statistical physics to study complex systems, broadly defined as systems with many interacting entities. For us, statistical physics is (non-rigorous) probability in high dimensions.

2.1 Microcanonical Ensemble

How can we move from Newton's laws of mechanics to an equilibrium statistical description where time is not involved? Considering a system of N particles, in principle, in order to describe the system, one would have to track the dynamics of all particles. According to Newton's laws, the dynamics of a system of particles is described by $6N$ coupled differential equations:

$$\begin{aligned} \frac{\mathrm{d}\boldsymbol{x}_i}{\mathrm{d}t} &= \boldsymbol{v}_i \\ \frac{d\boldsymbol{v}_i}{dt} &= \frac{1}{m}\boldsymbol{f}_i(\boldsymbol{x}_1, ..., \boldsymbol{x}_N, \boldsymbol{v}_1, ..., \boldsymbol{v}_N) \end{aligned} \tag{1}$$

Here $\boldsymbol{x}_i \in \mathbb{R}^3$ and $\boldsymbol{v}_i \in \mathbb{R}^3$ are the position and velocity of the i-th particle, m is the mass of the particle, and $\boldsymbol{f}_i$ is the force acting on the i-th particle. All quantities are time-dependent, although sometimes for conciseness we won't write the dependence explicitly as in $\boldsymbol{x}_i(t)$.

Once an initial condition, $\boldsymbol{x}_i(t=0), \boldsymbol{v}_i(t=0)$, $i = 1, ..., N$, is given, the system is completely determined at all times.

In practice, integrating the equations, using some numerical method such as Euler or Runge-Kutta, is not feasible for large N. Recall that the number of particles in one mole of a substance is the Avogadro number, $N_A \approxeq 10^{23}$.

But it is thanks to the large system size that we can have a statistical description of the system.

We assume that there exists a function that we call *energy*, $E\Big(\{\boldsymbol{x}_i, \boldsymbol{v}_i\}_i\Big)$, that is preserved in time by the dynamical system (this is true for the so-called Hamiltonian systems). We call E_0 the initial energy value, $E_0 = E\Big(\{\boldsymbol{x}_i(t=0), \boldsymbol{v}_i(t=0)\}_i\Big)$.

Boltzmann's intuition is that the system will explore all the microscopic configurations that are accessible to it, i.e. all the configurations $\{\boldsymbol{x}_i, \boldsymbol{v}_i\}_i$ that have energy E_0. These configurations are completely equivalent from the perspective of an observer measuring the system's macroscopic properties and for whom the microscopic details are irrelevant. After some time, which we call "equilibration time", the system will forget the specific initial condition it started with. For all practical purposes, a configuration observed at large times will be statistically equivalent to a configuration sampled from the uniform distribution over all configurations with energy E_0.

In order to formalize this concept, we introduce the statistical distribution known as *microcanonical ensemble*. Given a configuration C (in our gas case $C = \{\boldsymbol{x}_i, \boldsymbol{v}_i\}_{i=1}^N$), the probability of finding the system in that configuration is given by

$$p_{\text{micro}}(C) = \frac{\delta(E(C) - E_0)}{\Omega_{\text{micro}}(E_0)} \tag{2}$$

Ω is called the microcanonical partition function, guaranteeing the correct normalization of $p_{\text{micro}}(C)$. It just counts the number of accessible configurations:

$$\Omega_{\text{micro}}(E_0) = \sum_C \delta(E(C) - E_0)) \tag{3}$$

The sum is over the whole configuration space and should be replaced with an integral in the case of continuous variables. The notation $\delta(x)$ has to be interpreted as follows:

- If x can take only discrete values, $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ if $x \neq 0$. This is known as the *Kronecker delta*.
- If x is continuous, $\delta(x) = 0$ if $x \neq 0$ and $\delta(0) = +\infty$, s.t. $\int \mathrm{d}x\, \delta(x) = 1$. This is known as the *Dirac delta*.

Now we make a very strong assumption, which is called the *ergodicity* assumption. We assume that during the dynamics, at large times the system explores all the accessible configurations and spends the same amount of time in each of them.

Therefore, any quantity of interest, that we call an *observables*, that we would normally compute by averaging over time, we can compute by averaging over the microcanonical ensemble. That is, we can write:

$$\langle \mathcal{O} \rangle = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathrm{d}t'\, \mathcal{O}(C_{t'}) = \sum_C p_{\text{micro}}(C)\, \mathcal{O}(C) \tag{4}$$

Here $\mathcal{O}$ is an observable, e.g. the energy $E(C)$, and $\langle \bullet \rangle$ denotes equivalently a time average or an ensemble average.

The microcanonical *entropy* is defined as $S_{\text{micro}}(E_0) = \log \Omega_{\text{micro}}(E_0)$. It is the log number of accessible configurations. Since Ω_{micro} is typically exponential in the N, the entropy is *extensive*, i.e. it scales with the size of the system as $O(N)$. Also, the energy $E(C)$ is typically an extensive quantity.

We can define the corresponding *intensive* quantities by dividing by the number of particles N. For example, we can define the average intensive energy as $e = \frac{\langle E \rangle}{N}$. Intensive quantities have a well-defined limit in the *thermodynamic limit* $N \to +\infty$.

2.2 Canonical Ensemble

The microcanonical ensemble is useful when the energy of the system is fixed. If a system is in contact with a reservoir, it can exchange energy with it, therefore its energy is not fixed. On the other hand, the reservoir has the role of keeping the temperature T of the system fixed. The energy of a system can also oscillate if there is some source of noise in the dynamics (see Langevin equation).

In this case, the statistical description of the system is given by the *Canonical Ensemble*:

$$p(C) = \frac{e^{-\beta E(C)}}{Z(\beta)} \tag{5}$$

where $\beta = \frac{1}{T}$ is the (adimensional) inverse temperature. It can be obtained from the microcanonical ensemble by considering a reservoir and expanding $S(E_{\text{tot}}) = S(E + E_{\text{res}})$ for small E.

The normalization factor Z is known as *partition function*:

$$Z(\beta) = \sum_C e^{-\beta E(C)} \tag{6}$$

We will show that it plays a big role in the theory. In fact, the partition function contains all the relevant information about the system. In terms of the partition function, we can write the average energy as

$$\langle E \rangle = -\frac{\partial}{\partial \beta} \log Z(\beta). \tag{7}$$

The logarithm of the partition function is called the *free energy* of the system. We denote it in its intensive version as f:

$$f_N(\beta) = -\frac{1}{\beta N} \log Z(\beta). \tag{8}$$

Since Z is typically exponential in N, the free energy is an intensive quantity.

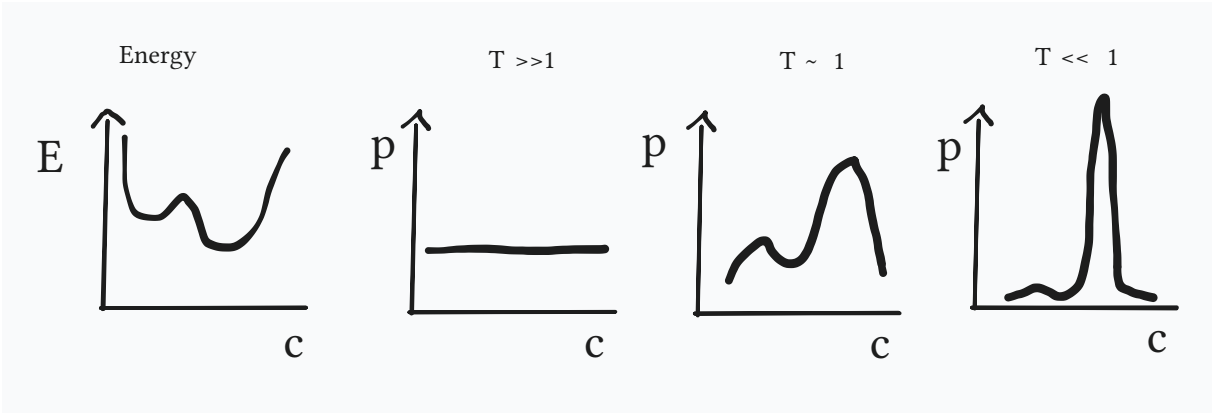

Figure 1: A pictorial representation of an energy function (left) and the corresponding Boltzmann distribution at high (center left), intermediate (center right), and small temperature (right).

Notice the dependence of $p(C)$ on the inverse temperature β:

- in the "high temperature" limit, $\beta \to 0$, the distribution becomes flat over all configurations. In the partition function Equation 6 al configurations contribute equally.
- in the "low temperature" limit, $\beta \to \infty$, the distribution is peaked around the minimum energy configuration (recall that $E(C)$ in). The partition function is dominated by the configurations with minimum energy.

See Figure 1 for a pictorial representation of the Boltzmann distribution at different temperatures.

The parameter β trades off the energy of the system with the entropy of the system.

In fact, let's define the intensive *entropy* $s(e)$ for fixed intensive energy e as

$$s(e) = \frac{1}{N} \log \sum_C \delta\left(e - \frac{E(C)}{N}\right). \tag{9}$$

We can rewrite the partition function as

$$\begin{aligned} Z(\beta) &= \sum_C e^{-\beta E(C)} = \sum_C e^{-\beta E(C)} \int \mathrm{d}e\, \delta\left(e - \frac{E(C)}{N}\right) \\ &= \int \mathrm{d}e\, e^{N(s(e)-\beta e)} \approx e^{N(s(e_*(\beta))-\beta e_*(\beta))} \end{aligned} \tag{10}$$

Where in the last step we used the Laplace approximation (discussed later), valid in the thermodynamic limit (i.e. $N \to +\infty$): the dominant contribution to the partition function is given by the configurations with energy given by

$$e_*(\beta) = \operatorname*{argmax}_e s(e) - \beta e. \tag{11}$$

For β small, e_* corresponds to maximum entropy (and typically high energy) configurations, while for β large, e_* corresponds to minimum energy configurations (with typically low corresponding entropy).

Let's make two further remarks:

- In the thermodynamic limit, the average value of the Energy corresponds to e_*:

$$\lim_{N \to +\infty} \frac{1}{N} \langle E \rangle = e_*(\beta) \tag{12}$$

- Again in the thermodynamic limit, the free energy of a system can be decomposed in an energetic and an entropic one as follows:

$$f(\beta) = \lim_{N \to \infty} f_N(\beta) = e_*(\beta) - \frac{1}{\beta} s(e_*(\beta)). \tag{13}$$

3 The Curie-Weiss Model

3.1 Model Definition

Let's consider a system of N binary variables, also known as Ising spins, each of which can take one out of two values, −1 and +1. A configuration is denoted by $\boldsymbol{\sigma} \in \{-1, +1\}^N$.

The energy of the system is given by

$$E(\boldsymbol{\sigma}) = -\frac{J}{N} \sum_{i<j} \sigma_i \sigma_j - h \sum_i \sigma_i, \tag{14}$$

where $J > 0$ is called the coupling constant and h is the external field. The coupling is divided by N to keep the energy extensive (i.e. $O(N)$).

The Ising model is the simplest model of a magnetic system. Intuitively, the spins tend to align with each other, i.e. to have the same value. The coupling constant J is positive, so the energy is minimized when the spins are aligned. They also tend to align with the external field h. See Figure 2 for a pictorial representation.

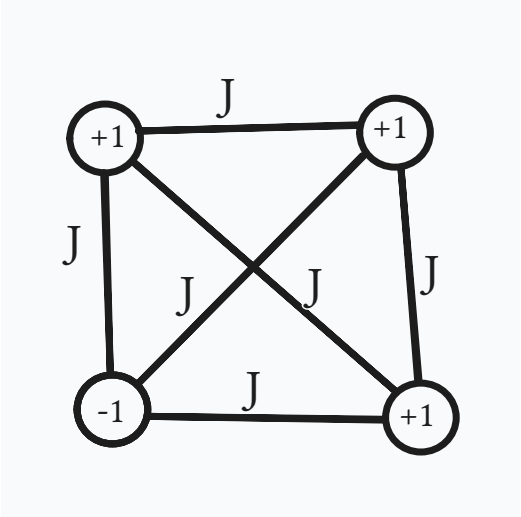

Figure 2: A Curie-Weiss model with $N = 4$ represented as a graph. The coupling J is a parameter associated with edges, and the spin variables σ_i are associated with nodes.

It is a mean-field model, i.e. it does not rely on any spatial structure: in the Curie-Weiss model, we assume that the spins are all-to-all connected. In other words, it is defined on a fully connected graph. Generally, the Ising model can be defined on an arbitrary graph, e.g. a d-dimensional lattice.

The partition function of the Curie-Weiss model is given by

$$Z = \sum_{\boldsymbol{\sigma}} e^{\beta \frac{J}{N} \sum_{i<j} \sigma_i \sigma_j + \beta h \sum_i \sigma_i}. \tag{15}$$

The quantity of interest we will look at is the ensemble average of the intensive magnetization of the system, $m = \frac{1}{N} \langle \sum_i \sigma_i \rangle$. Notice that we have the following relations:

$$m = \frac{1}{\beta N} \partial_h \log Z = -\partial_h f \tag{16}$$

In the last equality, we used the definition of the free energy $f = -\frac{1}{\beta N} \log Z$. In both Z and h we omitted the β, J, h, and N dependence for convenience.

The following sections are devoted to the computation of the partition function, free energy, and magnetization of the Curie-Weiss model.

3.2 The Case $J = 0$

When $J = 0$, the spins do not interact with each other. The partition function is factorized and we can compute it explicitly:

$$\begin{aligned} Z(J=0) &= \sum_{\boldsymbol{\sigma}} e^{\beta h \sum_i \sigma_i} \\ &= \left(\sum_{\sigma_1} e^{\beta h \sigma_1}\right)\left(\sum_{\sigma_2} e^{\beta h \sigma_2}\right) \dots \left(\sum_{\sigma_N} e^{\beta h \sigma_N}\right) \\ &= (2\cosh(\beta h))^N \end{aligned} \tag{17}$$

Correspondingly, also the probability distribution is factorized:

$$p(\boldsymbol{\sigma}) = \prod_{i=1}^{N} \frac{e^{\beta h \sigma_i}}{2\cosh(\beta h)} = \prod_{i=1}^{N} p(\sigma_i) \tag{18}$$

The magnetization of any variable is given by $m_i = \langle \sigma_i \rangle = \tanh(\beta h)$.

3.3 A Combinatorial Approach

Now let's go back to the general case $J \geq 0$. First of all, we show that the energy depends on the configuration only through the magnetization $M = \sum_i \sigma_i$. In fact, we have

$$\begin{aligned} E(\boldsymbol{\sigma}) &= -\frac{J}{N} \sum_{i<j} \sigma_i \sigma_j - h \sum_i \sigma_i \\ &= -\frac{J}{2N} \sum_{i \neq j} \sigma_i \sigma_j - hM \\ &= -\frac{J}{2N} \sum_{i,j} \sigma_i \sigma_j + \frac{J}{2} - hM \\ &= -\frac{J}{2N} M^2 + \frac{J}{2} - hM. \end{aligned} \tag{19}$$

From now on will discard the term $\frac{J}{2}$ in the energy, since it is a constant that can be reabsorbed in the partition function and also it is subleading, $O(1)$, in N. Since the magnetization M takes values in $\{-N, -N+2, -N+4, \dots, N-2, N\}$, we can write the partition function as

$$Z = \sum_{M} \exp\left(\beta J \frac{M^2}{2N} + \beta h M\right) \mathcal{N}(M), \tag{20}$$

where we defined the function $\mathcal{N}(M)$ that counts the number of configurations with magnetization M:

$$\mathcal{N}(M) = \sum_{\boldsymbol{\sigma}} \delta\left(M - \sum_i \sigma_i\right) \tag{21}$$

Using basic facts in combinatorics, one can see that $\mathcal{N}(M)$ is given by the binomial coefficient:

$$\mathcal{N}(M) = \binom{N}{\frac{N+M}{2}} = \frac{N!}{\left(\frac{N+M}{2}\right)!\left(\frac{N-M}{2}\right)!}, \tag{22}$$

where $\frac{N+M}{2}$ is the number of spins "up".

Use the Stirling approximation $N! \approx \sqrt{2\pi N} N^N e^{-N}$, after some work one is able to show that

$$\mathcal{N}(M) \approx e^{N H_2(\frac{M}{N})}. \tag{23}$$

where $H_2(x) = -\frac{1+x}{2}\log\frac{1+x}{2} - \frac{1-x}{2}\log\frac{1-x}{2}$ is the binary entropy function.

Exercise - Binary Entropy
Show that $\mathcal{N}(M) \approx e^{N H_2(\frac{M}{N})}$.

Now we can write the partition function as

$$\begin{aligned} Z &\approx \sum_M \exp\left(\beta J \frac{M^2}{2N} + \beta h M + N H_2\left(\frac{M}{N}\right)\right) \\ &\approx \int_{-1}^{1} \mathrm{d}m \exp\left[N\left(\frac{1}{2}\beta J m^2 + \beta h m + H_2(m)\right)\right]. \end{aligned} \tag{24}$$

In the last line, we used the large N limit to rewrite the sum as an integral over the auxiliary variable $m = \frac{M}{N}$. The integral can now be evaluated using the Laplace method.

For the time being, we stop here, and we take a different and more generic route to perform the computation, one that doesn't involve dealing with combinatorics.

3.4 The Field-Theoretical Approach

Let's tackle the computation of the partition function for the Curie-Weiss model using what is called a field-theoretical approach, which is a standard technique in statistical physics.

$$\begin{aligned} Z &= \sum_\sigma e^{-\beta E(\sigma)} = \sum_\sigma e^{\beta\frac{J}{N}\sum_{i<j}\sigma_i\sigma_j + \beta h\sum_i \sigma_i} \\ &= \sum_\sigma e^{\frac{1}{2}\beta\frac{J}{N}\sum_{i,j}\sigma_i\sigma_j + \beta h\sum_i\sigma_i - \frac{1}{2}\beta J} \\ &= \sum_\sigma e^{\frac{1}{2}\beta\frac{J}{N}\left(\sum_i\sigma_i\right)^2 + \beta h\sum_i\sigma_i - \frac{1}{2}\beta J} \\ &= \sum_\sigma \int_{-\infty}^{+\infty} \frac{\mathrm{d}\psi}{\sqrt{2\pi/\beta J N}} \exp\left(-\frac{1}{2}\beta J N\psi^2 + \beta J\psi\sum_i\sigma_i + \beta h\sum_i\sigma_i - \frac{1}{2}\beta J\right) \end{aligned} \tag{25}$$

In the last line, we applied what is called the Hubbard-Stratonovich transformation:

$$e^{\frac{1}{2}\frac{b^2}{a}} = \int_{-\infty}^{+\infty} \frac{\mathrm{d}x}{\sqrt{2\pi/a}} e^{-\frac{1}{2}ax^2 + bx} \tag{26}$$

The Hubbard-Stratonovich is just a Gaussian integral done in reverse order! It is a very useful trick in statistical physics, used to decouple interacting terms into a sum of independent terms coupled to a single auxiliary variable. We will see use it many times in the course. There is a degree of freedom in choosing the terms a and b in the transformation. In Equation 25 we chose them in such a way that we will be able to identify ψ with the magnetization of the system, as we will show later.

The nice thing in the last line of Equation 25 is that everything is factorized over the site indices i. We paid the price of introducing a new variable ψ. Sometimes we will call it a *field*.

Notice that in the exponent most terms are $O(N)$. Since we are interested in the *thermodynamic limit* $N \to \infty$, we can ignore subleading terms in the partition function.

$$
\begin{aligned}
Z &\approx \sum_{\sigma\in\{-1,+1\}^N} \int_{-\infty}^{+\infty} \mathrm{d}\psi\, e^{-\frac{1}{2}\beta J N\psi^2 + \beta J\psi\sum_i \sigma_i + \beta h\sum_i \sigma_i} \\
&= \int_{-\infty}^{+\infty} \mathrm{d}\psi\, e^{-\frac{1}{2}N\beta J\psi^2} \left(\sum_{\sigma\in\{-1,+1\}} e^{\beta J\psi\sigma + \beta h\sigma}\right)^N \\
&= \int_{-\infty}^{+\infty} \mathrm{d}\psi\, e^{-\frac{1}{2}\beta J N\psi^2 + N\log(2\cosh(\beta J\psi + \beta h))}
\end{aligned}
\tag{27}
$$

Notice that in the exponent we now have a factor N multiplying an N independent term.

We can now use the Laplace method to evaluate the integral:

$$
\int \mathrm{d}x\, e^{Nf(x)} \approx e^{Nf(x_*)} \sqrt{\frac{2\pi}{-Nf''(x_*)}} \quad \text{(for large } N\text{)} \tag{28}
$$

Here x_* is the maximum of $f(x)$.

Exercise - Laplace approximation
Derive the previous formula by Taylor expanding $f(x)$ around its maximum.

In our case, by computing the saddle point of the exponent, we have

$$
\psi_* = \tanh(\beta J\psi_* + \beta h) \tag{29}
$$

This is a self-consistent equation for ψ^* that has to be solved numerically by iteration.

Finally, we can write the free energy of the problem as

$$
\begin{aligned}
f_\beta(J,h) &= \lim_{N\to\infty} -\frac{1}{\beta N}\log Z \\
&= \frac{1}{2}J\psi_*^2 - \frac{1}{\beta}\log(2\cosh(\beta J\psi_* + \beta h))
\end{aligned}
\tag{30}
$$

Notice that ψ_* implicitly depends on β, J and h.

Let's compute the magnetization of the system. It can be obtained by taking the derivative of the free energy with respect to h:

$$
\partial_h f_\beta(J,h) = -\lim_{N\to+\infty} \frac{1}{N}\langle\sum_i \sigma_i\rangle = -m. \tag{31}
$$

Using the saddle point expression for f we have

$$
\begin{aligned}
m &= -\partial_h f_\beta(J,h) \\
&= \tanh(\beta J\psi_* + \beta h) \\
&= \psi_*
\end{aligned}
\tag{32}
$$

Therefore, the field ψ at the saddle point corresponds to the magnetization! Therefore, we can just rename ψ to m and express everything in terms of the magnetization:

$$
f_\beta(J,h) = \frac{1}{2}Jm^2 - \frac{1}{\beta}\log(2\cosh(\beta h + \beta J m)), \tag{33}
$$

with

$$m = \tanh(\beta(h + Jm)). \tag{34}$$

Notice that in the absence of coupling, $J = 0$, we have $m = \tanh(\beta h)$, which is consistent with our previous finding about the magnetization of a single spin in an external field h. Equation 34 can be interpreted as the equation governing a single "effective particle", representing the average behavior of the system, where Jm is the effective field felt by the particle due to the interaction with the other spins. This explains the term "mean-field model".

Equation 33 along with Equation 34 are the final expressions for the free energy of the Curie-Weiss model.

Remark I: We have reduced the problem of computing a very high-dimensional distribution over 2^N configurations, to the problem of evaluating a simple self-consistent equation for a scalar variable m, the magnetization.

Remark II: We somewhat abused the notation by calling m both an auxiliary variable that we introduced while doing formal manipulations of the partition function, and the (intensive) magnetization observable defined as $m = \frac{1}{N}\langle \sum_i \sigma_i \rangle$. In the thermodynamic limit, the two quantities coincide.

Remark III: The free energy f, seen as a function of m and fixing the other parameters, can be seen as a variational expression for the true free energy of the system. This means that

$$\lim_{N\to+\infty} -\frac{1}{\beta N} \log Z = \min_m f(m) \tag{35}$$

The self-consistent equation Equation 34 can be obtained as the saddle point equation of the variational free energy, i.e. by setting $\partial_m f(m) = 0$.

3.5 Phase Transitions

3.5.1 Phase transition in temperature ($h = 0$)

Let's analyze the solution of Equation 34 for different values of β, assuming $h = 0$. We also set $J = 1$ without loss of generality, since we can always reabsorb it in the definition of β.

One can solve the equation numerically by iteration as follows. Start from an initial guess m_0, and for each iteration $t > 0$ compute $m_{t+1} = \tanh(\beta m_t)$ until convergence. The solution is a fixed point of the function $m = \tanh(\beta m)$. We plot the solution in Figure 3. See the notebook `curie-weiss.ipynb` for the implementation.

Notice that there is a critical value $\beta_c = 1$ such that for $\beta < 1$ there is only one solution, while for $\beta > 1$ there are three solutions. The solution $m = 0$ is always present, but for $\beta > 1$ there are two additional and opposite solutions, $\pm m$. The system is said to undergo a *phase transition* at $\beta = \beta_c$. Let's call $m_*(\beta)$ the upper branch of the solution.

From the Figure 3 (Right) one can see that m_* is a continuous function of β, but it is not differentiable at $\beta = 1$. Therefore the phase transition is called a *continuous* or a *second order*. m is called the order parameter of the transition. That is, it is a quantity that is zero in one phase and different from zero in the other phase. The two phases are called the *paramagnetic* phase, where $m = 0$, and the *ferromagnetic* phase, where $m \neq 0$.

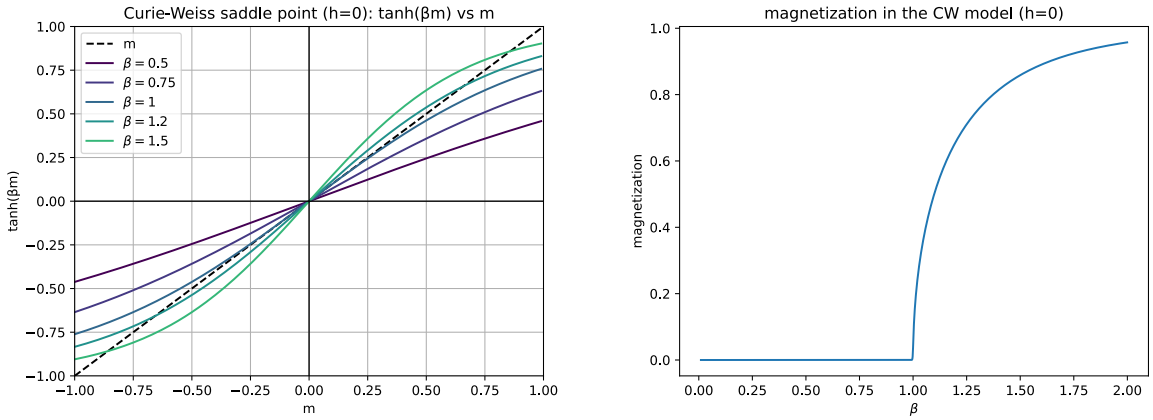

Figure 3: (Left) r.h.s. of the equation $m = \tanh(\beta m)$ for different values of β. The values of m that solve the equation are the points of the curve intersecting the bisector. For $\beta < 1$ there is only one solution, while for $\beta > 1$ there are three solutions. (Right) The positive solution as a function of the inverse temperature β.

The value $\beta_c = 1$ is called the *critical point* of the transition. Its value can be computed by noticing in Figure 3 (Left) that one starts having multiple solutions when the slope of the curve at $m = 0$ is equal to 1. Since at $\tanh(\beta m) = \beta m + o(m)$ for small m, we have that the slope is equal to 1 when $\beta = \beta_c = 1$.

It is also instructive to look at the variational free energy from Equation 33 as a function of m for different values of β. We show the result in Figure 4. Remember that the stationary points of the free energy are the solutions of the self-consistent equation Equation 34: in the paramagnetic phase $f(m)$ has a unique minimum at $m = 0$, while in the ferromagnetic phase the stationary points of $f(m)$ are three.

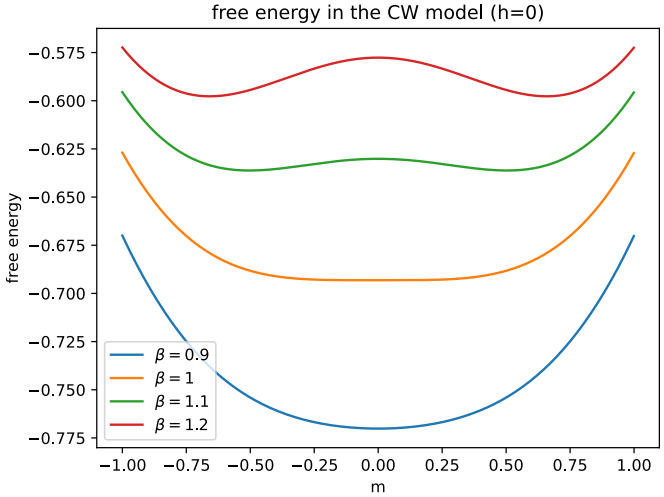

Figure 4: Variational free energy in the CW model as a function of the magnetization with zero external magnetic field.

3.5.2 First order phase transition

Consider now the case $h \neq 0$, setting a value of $\beta > 1$. As shown in Figure 5 (Left), the free energy has two minima when $h \neq 0$. When $h > 0$ the global minima of the free energy is for $m > 0$ whereas if $h < 0$ it has negative magnetization $m < 0$.

We can again keep track of the location of the global minima of the free energy as a function of h (see the notebook `curie-weiss.ipynb` for the implementation). Let's call it $m_*(h)$. As shown in the right panel of Figure 5, the global minima $m_*(h)$ has a discontinuity going from $h < 0$ to $h > 0$ at $h = 0$. A discontinuity in the order parameter is the signature of a *first order* phase transition. Notice that this first order phase transition is only present if $\beta > 1$.

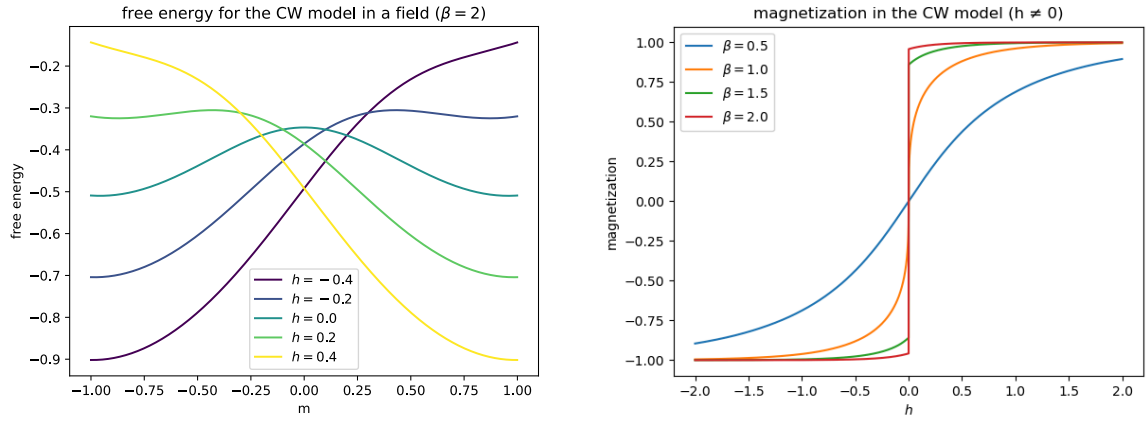

Figure 5: (Left) Variational free energy in the CW model as a function of the magnetization for different values of the external field and $\beta = 2$.

4 The p-spin Ferromagnetic Model

WARNING: Section 4 is not a mandatory read for the exam, but you may want to go over it for a better understanding of the field-theoretical techniques.

4.1 Ising p-Spin

Let's now consider a more complex model, the p-spin ferromagnetic model. We work again with Ising spins, $\sigma_i \in \{-1, +1\}$, but now the interactions are between p spins.

The energy of the system is given by

$$E(\boldsymbol{\sigma}) = -\frac{1}{2}\frac{Jp!}{N^{p-1}} \sum_{i_1<\ldots<i_p} \sigma_{i_1}\sigma_{i_2}\ldots\sigma_{i_p} \tag{36}$$

where $J_{i_1,\ldots,i_p}$ are the coupling constants. The summation is over all possible choices of p indexes in the set $1, \ldots, N$, without repetitions and ordering. Since we have $\binom{N}{p} \approx \frac{N^p}{p!}$ terms in the sum, we divide by N^{p-1} to keep the energy extensive (i.e. $O(N)$). The model is called p-spin because the interactions involve p spins.

Let's try to compute the partition function.

$$\begin{aligned} Z &= \sum_\sigma e^{-\beta E(\sigma)} = \sum_\sigma e^{\frac{1}{2}\frac{\beta Jp!}{N^{p-1}} \sum_{i_1<\ldots<i_p} \sigma_{i_1}\ldots\sigma_{i_p}} \\ &\approx \sum_\sigma e^{\frac{1}{2}\beta JN\left(\frac{\sum_i \sigma_i}{N}\right)^p} \end{aligned} \tag{37}$$

In the last line, we symmetrized the sum, ignoring subleading contributions in N from diagonal terms.

Now we introduce the integral representation of the Dirac delta function:

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \mathrm{d}\hat{x}\, e^{-i\hat{x}x} \tag{38}$$

We can now write the partition function as

$$\begin{aligned} Z &= \sum_{\sigma} \int \mathrm{d}m\, \delta\left(m - \frac{1}{N}\sum_i \sigma_i\right) e^{\frac{1}{2}\beta J N m^p} \\ &= \sum_{\sigma} \int \frac{\mathrm{d}m\, \mathrm{d}\hat{m}}{2\pi/N}\, e^{\frac{1}{2}\beta J N m^p - i\hat{m}\left(Nm - \sum_i \sigma_i\right)} \\ &= \int \frac{\mathrm{d}m\, \mathrm{d}\hat{m}}{2\pi/N}\, e^{N\varphi(m,\hat{m})} \end{aligned} \tag{39}$$

with

$$\varphi(m,\hat{m}) = \frac{1}{2}\beta J m^p - i\hat{m}m + \log(2\cosh(i\hat{m})) \tag{40}$$

We are now ready to evaluate the partition function using the saddle point method, which is an extension of the Laplace method to the complex plane.

The saddle point equations are given by

$$\partial_m \varphi(m,\hat{m}) = 0, \quad \partial_{\hat{m}} \varphi(m,\hat{m}) = 0. \tag{41}$$

The first gives

$$i\hat{m} = \frac{p}{2}\beta J m^{p-1}. \tag{42}$$

We see here that $i\hat{m}$ evaluated at saddle point is a real number.

The second equation gives

$$m = \tanh(i\hat{m}). \tag{43}$$

We can use the first SP to write everything as a function of m:

$$\begin{aligned} \varphi(m) &= -\frac{p-1}{2}\beta J m^p + \log\left(2\cosh\left(\frac{p}{2}\beta J m^{p-1}\right)\right), \\ m &= \tanh\left(\frac{p}{2}\beta J m^{p-1}\right). \end{aligned} \tag{44}$$

Exercise - Ising P-Spin
Solve the saddle point equations Equation 44 for $p = 3$. Is there a transition? If yes, what kind of transition is it (first or second order)?

4.2 Spherical p-Spin

We consider the model with p-wise interaction but now with continuous variables instead of Ising ones. We denote them with $\boldsymbol{x} \in \mathbb{R}^N$, they are subject to the spherical constraint $\|\boldsymbol{x}\|^2 = N$. Notice that the same holds in the model with Ising spins: $\sum_i \sigma_i^2 = N$: the spherical variable $\boldsymbol{x}$ lives on the surface of the hypersphere of radius $\sqrt{N}$ circumscribing the hypercube defined by the σ_i. This is called the *spherical p-spin model*:

$$E(\boldsymbol{x}) = -\frac{1}{2}\frac{Jp!}{N^{p-1}} \sum_{i_1 < \ldots < i_p} x_{i_1} x_{i_2} \ldots x_{i_p} \tag{45}$$

The partition function can be written as follows:

$$Z = \int \mathrm{d}\boldsymbol{x}\, \delta\left(N - \sum_i x_i^2\right) e^{\frac{1}{2}\frac{\beta J p!}{N^{p-1}} \sum_{i_1<\ldots<i_p} x_{i_1} x_{i_2} \ldots x_{i_p}}. \tag{46}$$

The computation then proceeds as before, but now we have to deal with the spherical constraint.

$$\begin{aligned} Z &= \int \mathrm{d}\boldsymbol{x}\, \delta\left(N - \sum_i x_i^2\right) \int \mathrm{d}m\, \delta\left(m - \frac{1}{N}\sum_i x_i\right) e^{\frac{1}{2}\beta J N m^p} \\ &= \int \mathrm{d}\boldsymbol{x}\, \mathrm{d}\hat{x}\, \mathrm{d}m\, \mathrm{d}\hat{m}\, e^{\frac{1}{2} i\hat{x}\left(N - \sum_i x_i^2\right) + i\hat{m}\left(Nm - \sum_i x_i\right) + \frac{1}{2}\beta J N m^p} \\ &= \int \mathrm{d}\hat{x}\, \mathrm{d}m\, \mathrm{d}\hat{m}\, e^{N\left(i\hat{x} + i\hat{m}m + \frac{1}{2}\beta J m^p\right)} \left(\int \mathrm{d}x\, e^{-\frac{1}{2} i\hat{x}x^2 - i\hat{m}x}\right)^N \\ &= \int \mathrm{d}\hat{x}\, \mathrm{d}m\, \mathrm{d}\hat{m}\, e^{N\left(i\hat{x} + i\hat{m}m + \frac{1}{2}\beta J m^p\right)} \left(\int \mathrm{d}x\, e^{-\frac{1}{2} i\hat{x}x^2 - i\hat{m}x}\right)^N \end{aligned} \tag{47}$$

We can assume that at saddle point $i\hat{x}$ and $i\hat{m}$ will be real numbers, as we will self-consistently check later.

Exercise - Spherical p-Spin
Continue the calculation for the spherical p-spin model, deriving the saddle point equations.

5 Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms used to sample from complex probability distributions when direct sampling is difficult. These methods simulate a simple stochastic process, namely a Markov chain, whose stationary distribution is the desired target distribution, and therefore generate the desired samples by simulating the chain.

MCMC methods, and Metropolis-Hastings among them, provide a powerful framework for sampling from complex distributions. By ensuring the property of detailed balance, they guarantee convergence to the desired distribution, making them essential tools in statistical physics, Bayesian inference, and machine learning.

5.1 Markov Processes

A discrete-time *Markov process* (also called a Markov chain) is a stochastic process $(X_0, X_1, ...)$ where the probability of the next state depends only on the current state and not on the past states.

This means that we can write the probability of a given trajectory using conditional probabilities in the form

$$p(x_0, x_1, ..., x_T) = p_0(x_0) \prod_{t=1}^{T} p_t(x_t \mid x_{t-1}). \tag{48}$$

The process is also called *stationary* if the conditional probabilities don't depend explicitly on time, $p_t(x_t \mid x_{t-1}) = p(x_t \mid x_{t-1})$. From now on we will always consider stationary processes.

We assume that the random variables X_t take value in a discrete set, the set of state S. Thanks to Equation 48, a stationary Markov process is completely characterized by the initial probability distribution $p(x_0)$ and by the so-called *transition matrix*

$$P_{xy} = \mathbb{P}(X_{t+1} = x \mid X_t = y). \tag{49}$$

P is a *stochastic* matrix, meaning that it satisfies

$$\sum_{x \in S} P_{xy} = 1 \quad \forall y \in S, \tag{50}$$

or in other words, starting from state s, the process will surely move somewhere in S. A Markov chain is said to be *irreducible* if it is possible to reach any state from any other state in a certain number of time steps, and *aperiodic* if it does not exhibit cyclic behavior. A Markov chain that is both irreducible and aperiodic is said to be *ergodic*.

5.2 Stationary Distribution

Call $p_t(x)$ the distribution of random variable X_t. Then, by definition of transition matrix, it holds that

$$p_t(x) = \sum_y P_{xy}\, p_{t-1}(y) \tag{51}$$

Notice that this construction can be iterated:

$$p_t(x) = \sum_y P^2_{xy}\, p_{t-2}(y). \tag{52}$$

The Markov chain is said to be *irreducible* if for any two states x and y, there exists an integer n such that $P^n_{xy} > 0$, that is, there always exists a finite probability path leading from any state to any other state. A Markov chain is said to be *invariant* with respect to a distribution π if

$$\pi(x) = \sum_y P_{xy}\,\pi(y), \tag{53}$$

that can be written as a vector equation as

$$\pi = P\pi. \tag{54}$$

The distribution π is said to be a *stationary distribution* of the Markov chain: it doesn't change after an iteration of the dynamics. From an algebraic perspective, π is a right eigenvector of P with eigenvalue 1.

We have the following two important facts:

- **First ergodic theorem:** consider a Markov chain that is irreducible and invariant with respect to a distribution π. Then π is the unique stationary distribution of the MC.
- **Second ergodic theorem:** If a Markov chain is ergodic and π-invariant, then

  $$\lim_{t\to+\infty} P^t_{xy} = \pi(x)\ \forall x, y \tag{55}$$

  and therefore we have convergence to the invariant distribution

  $$\lim_{t\to+\infty} p_t = \pi \tag{56}$$

  regardless of p_0

The time it takes to converge to the stationary distribution is called the *mixing time*.

Convergence is exponentially fast in time, with a rate that depends on λ_2 the second largest eigenvalue (in modulus) of P, by $\tau \simeq -\frac{1}{\ln(|\lambda_2|)}$. In fact it can be shown that

$$p_t(x) = \pi(x) + O\left(e^{-\frac{t}{\tau}}\right) \quad \text{for} \quad t \to +\infty. \tag{57}$$

These theorems can be proved by using the Perron-Frobenius theorem from linear algebra.

Exercise - Markov Chain Convergence
Prove equation Equation 57 by decomposing $p_0(x)$ as a linear combination of the right eigenvectors of P, then apply P for t times.

5.3 Detailed Balance Condition

A probability distribution $\pi(x)$ is said to satisfy the *detailed balance condition* with respect to a Markov chain if:

$$P_{yx}\pi(x) = P_{xy}\pi(y), \quad \forall x, y \in S. \tag{58}$$

This condition ensures that the Markov chain has π as its stationary distribution (to verify that, it is sufficient to sum over y on both sides of Equation 58), meaning that if the chain starts from π, it remains in π at all times. Notice that the inverse implication is not true: in general, π-invariance does not imply detail balance.

5.4 The Metropolis-Hastings Algorithm

The *Metropolis-Hastings (MH) algorithm* is a widely used MCMC method that constructs a Markov chain with a specified stationary distribution π. The algorithm proceeds as follows:

1. Start with an initial state X_0.
2. At each iteration $t < t_{\max}$:

- Propose a new state X' from a proposal distribution $q(X' \mid X_t)$.
- Compute the acceptance ratio:

$$A(X_t \to X') = \min\left(1, \frac{\pi(X')q(X_t \mid X')}{\pi(X_t)q(X' \mid X_t)}\right). \tag{59}$$

- Accept the new state with probability $A(X_t \to X')$; otherwise, remain at X_t.

3. Return the final configuration $X_{t_{\max}}$ as an approximate sample from π.

This process generates a Markov chain whose stationary distribution is π, provided that the chain is irreducible and aperiodic (i.e ergodic).

The only tunable part in the algorithm is the proposal distribution $q(X' \mid X_t)$.

Notice that the acceptance ratio can be evaluated even if the distribution π is known only up to a normalization constant, as it is often the case in statistical physics and in Bayesian statistics. This is a key feature of the Metropolis-Hastings algorithm.

The MCMC methods are widely used in statistical physics, machine learning, and statistics. They are general-purpose and very effective methods to sample from high-dimensional distributions.

5.5 Detailed Balance for MH

To show that Metropolis-Hastings satisfies detailed balance, we check that:

$$\pi(X_t)P(X_t \to X') = \pi(X')P(X' \to X_t). \tag{60}$$

Using the transition probability:

$$P(X_t \to X') = q(X' \mid X_t)A(X_t \to X'), \tag{61}$$

we obtain:

$$\pi(X_t)q(X' \mid X_t)A(X_t \to X') = \pi(X')q(X_t \mid X')A(X' \to X_t). \tag{62}$$

Substituting the acceptance ratio into the equation above, it follows that detailed balance holds, ensuring that $\pi(x)$ is the stationary distribution of the chain. Indeed, assume

$$\pi(X')q(X_t \mid X') > \pi(X_t)q(X' \mid X_t) \tag{63}$$

then,

$$A(X_t \to X') = \frac{\pi(X')q(X_t \mid X')}{\pi(X_t)q(X' \mid X_t)} \tag{64}$$

and

$$A(X' \to X_t) = 1 \tag{65}$$

Substituting these relations in Equation 62 yields the detailed balance condition. This implies that the MC constructed in the Metropolis-Hastings algorithm is π-invariant: the configurations reached at $t \to +\infty$ are distributed according to π.

5.6 MCMC for Curie-Weiss

The Curie-Weiss model is a simple example of a model that can be solved exactly with analytic computations. In general, this is not the case. In particular, the partition function is often intractable to compute, or can only be computed for large system size N. In order to sample from complex distributions, we can use MCMC methods.

The Metropolis-Hastings algorithm for a spin system with energy $E(\boldsymbol{\sigma})$ proceeds as follows:

1. Start from an initial configuration $\boldsymbol{\sigma}$.
2. Propose a new configuration $\boldsymbol{\sigma}'$ by flipping a random spin.
3. Compute the acceptance probability $A(\boldsymbol{\sigma} \to \boldsymbol{\sigma}') = \min\big(1, e^{-\beta(E(\boldsymbol{\sigma}')-E(\boldsymbol{\sigma}))}\big)$.
4. Accept the new configuration with probability equal to the acceptance probability.
5. Repeat from step 2.

Notice any move resulting in a decrease of the energy is accepted with probability 1.

In the notebook `curie-weiss.ipynb` we show how to implement the Metropolis-Hastings algorithm for the Curie-Weiss model.

5.7 Energy Based Models and Contrastive Divergence

In this section we want to show a popular application of Monte Carlo methods to a common task in machine learning, i.e. generative modeling.

We have seen a few examples of models in which an energy function $E(\boldsymbol{x})$, characterized by some parameters (e.g. the coupling J), defines a probability distribution over the configurations $\boldsymbol{x}$.

In machine learning, a common task called generative modeling consists of learning a distribution from a finite set of samples. Let's call $\mathcal{D} = \{\boldsymbol{x}^\mu\}_{\mu=1}^P$ the set of P samples that we have.

One may wonder how to learn the parameters of the energy function $E(\boldsymbol{x})$ from the samples, in order to be able to generate new samples from the same distribution.

Let's call E_θ our energy function, with θ the parameters that we want to learn from data. As a concrete example, in case we are working with ± 1 variables, a simple energy function is given by

$$E_\theta^{\text{ising}}(\boldsymbol{\sigma}) = -\sum_i h_i \sigma_i - \sum_{i<j} J_{i,j} \sigma_i \sigma_j. \tag{66}$$

and we call θ the collection of all parameters, $\theta = (J, h)$.

We write the corresponding Boltzmann distribution as

$$p_\theta(\boldsymbol{x}) = \frac{e^{-E_\theta(\boldsymbol{x})}}{Z_\theta}, \tag{67}$$

with

$$Z_\theta = \int \mathrm{d}\boldsymbol{x}\, e^{-E_\theta(\boldsymbol{x})}. \tag{68}$$

We remark that if the function family E_θ is expressive enough, the Boltzmann distribution can approximate any distribution.

The standard way to learn the parameters of a model is by log-likelihood maximization. We call $L(\theta)$ the loss function, given by the negative log-likelihood of the data:

$$L(\theta) = -\frac{1}{P}\sum_{\mu=1}^P \log p_\theta(\boldsymbol{x}^\mu) = -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \log p_\theta(\boldsymbol{x}^\mu) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\, E_\theta(\boldsymbol{x}) + \log Z_\theta. \tag{69}$$

If we want to minimize the loss / maximize the likelihood, we have to compute the gradient of the loss with respect to the parameters:

$$\nabla_\theta\, L(\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\, \nabla_\theta E_\theta(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}\sim p_\theta}\, \nabla_\theta E_\theta(\boldsymbol{x}). \tag{70}$$

We have written the gradient in terms of the expectations over samples from the dataset (sometimes called positive samples) and from the model itself (negative samples). While the first expectations is easy to evaluate, the second is computationally more demanding, since typically the partition function Z_θ is intractable to compute and the model is not easy to sample from.

The Contrastive Divergence (CD) algorithm, consists of sampling from the model distribution p_θ using MCMC methods, and approximating the true expectation with an empirical one using these samples. This algorithm has been very successful in training energy-based models, in particular in the context of Restricted Boltzmann Machines (RBMs), a type of energy-based model that was used in the '00s.

Thanks to his work on contrastive divergence and energy-based models, bridging machine learning and statistical physics, Geoffrey Hinton was awarded the Nobel prize for physics in 2024.

Exercise - Contrastive Divergence Ising
Derive the specific form of CD update rule for the Ising model, that is

$$\begin{aligned} J'_{ij} &= J_{ij} - \eta(\dots) \\ h'_i &= h_i - \eta(\dots) \end{aligned} \tag{71}$$

Now sample from an Ising (random J and h) model using Metropolis-Hasting. Train, using CD, another Ising model. Plot the MSE of J and h as a function of time with 1000 samples and $N = 10$.

Exercise - Binarized MNIST
Train the Ising model to learn the binarized MNIST dataset.

The dataset can be found here https://github.com/yburda/iwae/tree/master/datasets/BinaryMNIST

6 Random Graphs

6.1 Basic Graph Theory definitions

A few basic definitions from graph theory.

A *graph* $G = (V, E)$ is a pair composed of a set of vertices V and a set of edges E. We will assume nodes to be the integers $1, \dots, N$ and edges to be pairs of integers (i, j) with $i, j \in V$.

For *undirected* graphs, the edge (i, j) is equivalent to the edge (j, i). For *directed* graphs instead, the edge (i, j) is not equivalent to the edge (j, i). We will focus on undirected graphs.

We denote with $i \sim j$ the fact that i and j are connected by an edge, i.e. $(i, j) \in E$.

The sum over the edges of a graph is denoted as $\sum_{(i,j)\in E}$ or $\sum_{i\sim j}$ or just $\sum_{(i,j)}$.

The *neighborhood* of a node i is the set of nodes j such that $i \sim j$. We denote it as $\partial i = \{j \in V : j \sim i\}$.

The *degree* of a node i is the number of edges connected to it, i.e. $\deg_i = |\partial i|$.

The *adjacency matrix* A_{ij} is defined as $A_{ij} = 1$ if $i \sim j$ and $A_{ij} = 0$ otherwise.

A *path* $(i_0, i_1, \dots, i_\ell)$ of length ℓ is a sequence of nodes such that $i_k \sim i_{\{k+1\}}$ for $k = 0, \dots, \ell - 1$. A *loop* is a path such that $i_0 = i_\ell$. A *simple loop* is a loop such that no intermediate node is repeated.

A *connected component* is a subset $C \subseteq V$ such that there is a path connecting any two nodes in C. The nodes of a graph can be partitioned into disjoint connected components.

A graph is a *tree* if it has a single connected component and no loops. A graph is a *forest* if each connected component is a tree.

6.2 Notable Ensembles

A graph ensemble is a probability distribution over the space of graphs. A random graph is a graph sampled from a graph ensemble.

The following ensembles are particularly important:

- Erdos-Renyi ensemble: $\mathrm{ER}(N, p)$. A distribution over graphs with N nodes where a graph is sampled by including each edge independently with probability p. The expected number of edges is $\frac{N(N-1)}{2}p$ and the expected degree for any node is $c = (N-1)p$.
- Random Regular Graph Ensemble: $\mathrm{RRG}(N, c)$. This ensemble gives uniform probability over graphs with N node having all degree c. In order to sample, create c stubs for each node and pair them randomly.
- Configuration Model: $\mathrm{ConfModel}\big(N, \mu_{\mathrm{deg}}\big)$. Distribution over graphs with N node where the degrees are sampled independently from μ_{deg}. First sample the degrees, $d_i \sim \mu_{\mathrm{deg}}$ i.i.d., then create d_i stubs for each node and pair them randomly as in the RRG case.

6.3 Statistical Properties of ER Graphs

For simplicity, we will focus on Erdos-Renyi $G \sim \mathrm{ER}(N, p)$ with $p = \frac{c}{N}$. This scaling of p with N is important because it allows to keep the average degree constant while increasing the number of nodes N. Graphs from this distribution are sparse ($O(1)$ degree for each node) and have a number of edges $O(N)$.

Thanks to the fact that edges are included independently, it is easy to compute some basic properties for this ensemble.

For instance, the expected number of edges is

$$\mathbb{E}[\,|E|\,] = \mathbb{E}\left[\sum_{i<j} A_{ij}\right] = \frac{N(N-1)}{2}\mathbb{E}[A_{12}] = \frac{N(N-1)}{2}\frac{c}{N} = \frac{c(N-1)}{2} \approx \frac{cN}{2}, \tag{72}$$

where the last approximation holds for large N. The expected degree for any node i is

$$\mathbb{E}[\deg_i] = \mathbb{E}\left[\sum_{j;j\neq i} A_{ij}\right] = (N-1)\frac{c}{N} \approx c. \tag{73}$$

A more refined calculation gives the distribution of the degree d_i for a node i, which turns out to be Poisson distributed with average c for large N:

$$\begin{aligned} \mathbb{P}(\deg_i = d) &= \binom{N-1}{d}\Big(\frac{c}{N}\Big)^d\Big(1-\frac{c}{N}\Big)^{N-1-d} \\ &\approx \frac{N^d}{d!}\frac{c^d}{N^d}e^{-cN} \\ &= \frac{c^d e^{-cN}}{d!} = \mathrm{Poisson}(d; c). \end{aligned} \tag{74}$$

In the last equation we looked at a single node i, and we computed the degree probability across the realizations of the graph.

For the law of large numbers, when considering a single but large graph, and compute the frequency of nodes of degree d (the empirical degree distribution), we will obtain the same distribution.

Another interesting property of the $\mathrm{ER}(N, \frac{c}{N})$ ensemble is that graphs are locally tree-like. This means that if we look at a node i and its neighborhood at distance L (the neighbors, the neighbors of the neighbors, etc.), this subgraph looks like a tree. This is true with high probability (w.h.p., i.e. with

probability that goes to 1 across the realizations of the graph) for L that remains finite as N goes to infinity.

One can start from a root node i and look at the nodes at distance less or equal to L from it. At large N, this is given (in expectation) by $1 + c + c^2 + ... + c^L$ (show it!). This geometric progression converges to a finite value for $c < 1$ and diverges for $c > 1$.

In fact, for $c < 1$ the branching process that goes away from the root dies out and the expected number of nodes in the component of i is given by:

$$\mathbb{E}[\mathcal{C}_i] \approx 1 + c + c^2 + ... = \frac{1}{1-c}. \tag{75}$$

For $c < 1$ the connected component of any node i is a tree w.h.p. and has size $O(1)$.

For $c > 1$ the graph is still locally tree-like, but the branching process does not die out and, as we will show later, we will have a giant component. One can show that as long as a finite length L is kept, a loop involving i is rare by computing the probability of loop existence times the number of possible loops going through i. This is given by

$$\begin{aligned} \mathbb{P}(i \text{ in loop of length } L) &\approx \frac{(N-1)...(N-L+1)}{2} p^L \\ &\approx \frac{1}{N} \frac{c^L}{2} \to 0. \end{aligned} \tag{76}$$

In the entire graph on average we have only $\frac{c^L}{2L}$ loops of size L.

The typical size of loops is $L = O\Big(\frac{\log N}{\log c}\Big)$. It can be obtained by requiring that the number of nodes at distance L from a given one becomes $O(N)$, that is $c^L = O(N)$.

6.4 Giant Component

We want to describe the size of the largest component in a $\mathrm{ER}(N, \frac{c}{N})$ random graph in the thermodynamic limit $N \to +\infty$ and as a function of the average degree c.

It turns out that we have the following phenomenology:

- For $c > 1$, a finite fraction of the nodes belong to a single connected component. The other components are trees with finite size.
- For $c < 1$, any connect component is a tree with finite size w.h.p. (with high probability for large N).

Let's give a probabilistic argument to derive this result and more. In fact, we will derive a self-consistent equation for the size of the largest component (up to $O(N)$ order).

Call x_G the probability for a node to be a giant component.

$$x_G = \lim_{N\to\infty} \frac{\mathbb{E}[\,|C_{\max}|\,]}{N} = \mathbb{P}[i \ \text{ in giant component}]. \tag{77}$$

A node does not belong to the giant component if all its neighbors do not belong to the giant component. Therefore, if d is the (random) degree of the node, we can derive a self-consistent equation for x_G:

$$1 - x_G = \mathbb{E}_d (1 - x_G)^d. \tag{78}$$

Using the fact that d is Poisson distributed with average c, we obtain

$$1 - x_G = e^{-c x_G}. \tag{79}$$

This is a transcendental equation. For any c, it has the solution $x_G = 0$, which corresponds to the absence of a giant component. For $c > 1$ there is a second solution $x_G > 0$ which correctly describes the giant component.

Even though this is a very crude probabilistic argument that discards correlations among events, the solution from the resulting self-consistent equation has been proved to be correct.