APMLA L7: the Mean Field approach

Caterina De Bacco and Isabel Valera

1 Curie-Weiss ferromagnet

We start straight with a prototypical example of interacting system for which we want to perform a probabilistic interpretation: the Curie-Weiss model. It will be the first example of the Mean Field method, an approach needed to approximate a complicated joint distribution.

Consider a system made of N binary random variables (spins) σ_i, each one interacts with all the others in pairwise interactions of the same (small) magnitude $\frac{J}{N}$, i.e. a *complete* network. In addition, we have an external magnetic field acting, with the same magnitude h, on each of the single spins independently.

This is formalized as follows:

- $s_i \in \{\pm 1\}$, $i = 1, \cdots, N$: values of binary variables (i.e. a particular realization of σ_i);
- $\mathbf{s} = (s_1, \ldots, s_N)$: a N-dimensional realization of the spin-configuration;
- $\frac{J}{N} > 0$: pairwise couplings encoding pairwise interactions between the spin variables;
- h: local magnetic fields acting on individual spins.

A network representation is given in Fig. 1.

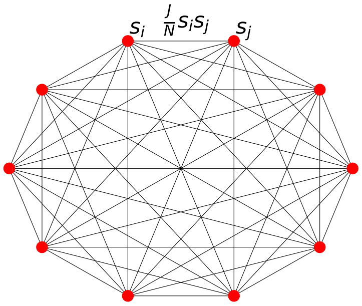

Figure 1: The complete network representation of a Curie-Weiss model. Each edge is an interaction of magnitude $\frac{J}{N} s_i s_j$. Here, interactions with the external field are not drawn.

In this system we want:

- Spins with the same value to have higher likelihood to interact.
- Spins "aligned" with the external field to have higher likelihood.

An example application could be opinion spreading, where $s_i = \pm 1$ represent one of two opposite opinions about a topic and people aligned are more likely to interact. In this case, h could be a global environmental effect, like a marketing campaign that makes people more likely to align to the advertisement message.

This can be formalized by considering an energy function (or Hamiltonian):

$$H(\boldsymbol{s}) = -\frac{J}{2N}\sum_{i,j} s_i s_j - h\sum_i s_i \quad , \tag{1}$$

which usually one aims at minimizing. This in fact can also be seen as a cost function in optimization problems. Indeed, the minimum (or ground state) of this function is when all spins are "aligned", i.e. they are all $+1$ or -1 (depending on the sign of h). In general, however, the system can be in a state away from the ground state, due to random fluctuations. Still, the higher the energy the less likely that configuration to be observed.
This can be set into a probabilistic framework by describing the system with a Boltzmann distribution:

$$P(\boldsymbol{s}) = \frac{e^{-\beta H(\boldsymbol{s})}}{Z} \tag{2}$$

$$Z = \sum_{\mathbf{s}} e^{-\beta H(\boldsymbol{s})} \quad , \tag{3}$$

where β is a parameter called "inverse-temperature" that shapes the distribution.

In words, the probability of observing a configuration $\mathbf{s}$ is given by $P(\boldsymbol{s})$ as above, a distribution shaped by the choice of Hamiltonian $H(\mathbf{s})$ and parameter β.

Obs1: in all the definitions above we omitted the explicit dependence on the parameters $\mathbf{J}, \mathbf{h}$. We keep them implicit because now we are focusing on the *forward* problem, i.e. the problem of estimating statistical observables (such as magnetization and correlations) under the Boltzmann distribution assuming the parameters $\mathbf{J}, \mathbf{h}$ as *given*. This means that we write $P(\boldsymbol{s})$, although to be explicit we would need to write $P(\boldsymbol{s}|\mathbf{J}, \mathbf{h})$.

Q1: how does the single variable s_i behave? We expect its behavior to depend on the state of the others.
Q2: what about the average $m \equiv \frac{1}{N}\sum_i s_i$? This is also called (instantaneous) *magnetization*. If everything behaves completely randomly, we expect this to average out to zero. Indeed, its behavior can be highly non trivial based on the values of the parameters.
Q3: similarly, what can we say about the pair correlations $\chi_{ij} \equiv s_i s_j$? Do they decay to zero on average?

In other words, we want to be able to compute marginals like:

$$P(s_i) = \sum_{s_j=\pm 1, s_j \in \mathbf{s}_{\setminus i}} P(\mathbf{s}) \tag{4}$$

$$P(s_i, s_j) = \sum_{s_k=\pm 1, s_k \in \mathbf{s}_{\setminus i,j}} P(\mathbf{s}) \quad , \tag{5}$$

where $\mathbf{s}_{\setminus i}$ is the set $\left\{s_j : j \neq i, \quad j = 1, \ldots, N\right\}$.

These are all relevant questions that one should ask in order to understand the system. In general, one is usually interested in estimating statistical observables $O(\mathbf{s})$. In particular one usually wants to study averages like:

$$\mathbb{E}[O] = \langle O \rangle = \sum_{\mathbf{s}} P(\mathbf{s}) O(\mathbf{s}) \quad , \tag{6}$$

where the expected value is under the Boltzmann distribution (2)[1].

[1] We omit the explicit distribution each time there is no confusion.

Few common examples are:

$$\begin{aligned}
\mathbb{E}[m] = \langle m \rangle &= \frac{1}{N}\sum_{\mathbf{s}} P(\mathbf{s}) \sum_i s_i && \text{expected average magnetization} \quad (7)\\
m_i \equiv \mathbb{E}[s_i] = \langle s_i \rangle &= \sum_{\mathbf{s}} P(\mathbf{s}) s_i = \sum_{s_i=\pm 1} P(s_i) s_i && \text{expected magnetization per spin} \quad (8)\\
\mathbb{E}\left[\chi_{ij}\right] = \langle s_i s_j \rangle &= \sum_{\mathbf{s}} P(\mathbf{s}) s_i s_j = \sum_{s_i, s_j=\pm 1} P(s_i, s_j) s_i s_j && \text{expected pair correlation btw } i, j \quad (9)
\end{aligned}$$

Obs1: notice the two equivalent notations $\mathbb{E}[m]$ and $\langle m \rangle$, in case the distribution with respect to which we are performing the average is obvious; otherwise we write $\mathbb{E}_Q[m]$ or $\langle m \rangle_Q$. We will use them interchangeably in the lectures.

1.1 Mean Field: physics intuition approach

We now want to answer to the questions above and solve the Curie-Weiss model. To do this, we focus on one individual i and write the Boltzmann weight in terms of s_i explicitly:

$$H(s) = -\sum_i s_i \left[\frac{J}{2N} \sum_{j \neq i} s_j + h \right] = -\sum_i s_i h_i(\mathbf{s}_{\setminus \mathbf{i}}) \quad . \tag{10}$$

This rewriting shows that s_i is affected by an interaction $h_i(\mathbf{s}_{\setminus \mathbf{i}})$, a combination of interactions with others and with the external field, as expected. However, $h_i(\mathbf{s}_{\setminus \mathbf{i}})$ is also a random variable, and due to the pairwise couplings one cannot separate the different contributions further in a trivial way.

Question: Can we at least approximate it with an easier expression?

Idea: let's assume that s_i feels an *average* contribution from the others, plus the external field (which is fixed as a parameter). This means to approximate:

$$\begin{aligned}
h_i(\mathbf{s}_{\setminus \mathbf{j}}) \approx \mathbb{E}_P\left[h_i(\mathbf{s}_{\setminus \mathbf{j}})\right] &= \mathbb{E}_P\left[\frac{J}{2N} \sum_{j \neq i} s_j \right] + h && (11)\\
&\approx \frac{J}{2} \langle m \rangle + h \equiv h^{MF} \quad . && (12)
\end{aligned}$$

where the last approximation is valid for large N, as the magnetization is not much impacted by a single variable, i.e. $m = \frac{1}{N}\sum_{j \neq i} s_j + \frac{s_i}{N} \rightarrow \frac{1}{N}\sum_{j \neq i} s_j$ for $N \rightarrow \infty$.

In other words, we assume that s_i is affected by a *mean field* h^{MF}.
Equivalently, we are ignoring fluctuations around the expected values. Then, we can approximate the full energy with:

$$H(s) \approx -\sum_i s_i h^{MF} \equiv H^{MF}(s) \quad , \tag{13}$$

which in turns allows us to obtain a nicer expression for the probability distribution:

$$P(\mathbf{s}) \approx P^{MF}(\mathbf{s}) = \frac{1}{Z^{MF}} \prod_i e^{s_i h^{MF}} \tag{14}$$

Question: is this helping us to calculate marginals like Eq. (4,5) ? **Yes**! Let's see how.

$$P_{MF}(s_i) = \frac{1}{Z^{MF}} e^{s_i h^{MF}} \prod_{j \neq i} \sum_{s_j=\pm 1} e^{s_j h^{MF}} = \frac{e^{s_i h^{MF}}}{2 \cosh h^{MF}} \quad , \tag{15}$$

where we used $Z^{MF} = \prod_j \sum_{s_j=\pm1} e^{s_j h^{MF}} = \left[2\cosh h^{MF}\right]^N$. This implies that:

$$\mathbb{E}_{MF}\left[s_i\right] = \sum_{s_i=\pm1} s_i P_{MF}(s_i) = \tanh h^{MF} = \tanh\left[\frac{J}{2}\langle m\rangle + h\right] \tag{16}$$

However, we are still missing an expression for $\langle m\rangle$. We can find it by closing it with what just learned and noticing that the RHS of (16) does not depend on i. In addition, $\mathbb{E}_{MF}\left[m\right] \equiv \langle m\rangle_{MF} = \frac{1}{N}\sum_i \mathbb{E}_{MF}\left[s_i\right] = \mathbb{E}_{MF}\left[s_i\right]$. If we finally impose that $\langle m\rangle_{MF} = \langle m\rangle$, i.e. we match the expected magnetization of the Mean Field approximation with that of the (hard to calculate) exact model we obtain:

$$\langle m\rangle_{MF} = \tanh\left[\frac{J}{2}\langle m\rangle_{MF} + h\right] \quad . \tag{17}$$

This is a self-consistency equation that can be solved by plotting the LHS and RHS and finding where they intersect, as a function of the parameters.
This will give the Mean Field (MF) solution: we can finally write the MF Hamiltonian using Eq. (11):

$$H^{MF}(\mathbf{s}) = -h^{MF}\sum_i s_i = -\left[\frac{J}{2}m^* + h\right]\sum_i s_i \quad , \tag{18}$$

where m^* is the solution of Eq. (17).
Question: is this a good approximation? In general **no**. However, for the Curie-Weiss model Mean Field becomes exact for $N \to \infty$.
Why? Because we have constant and small couplings J/N between variables, which make fluctuations small and variables weakly dependent.

Exercise: calculate $Var\left[h_i(\mathbf{s}_{\setminus \mathbf{i}})\right]$ and show that for the Curie-Weiss model this goes to 0 for large N. This is a non-rigorous way to show that the approximation (11) is indeed valid for this model.

In other more realistic scenarios, fluctuations matter and Mean Field does not work well anymore. However, there are several ways to improve it, as we will see in the next lectures.

2 Mean Field: Variational Approach

Abandoning the physics intuition, and wearing instead a general probabilistic hat, we could have approached the Curie-Weiss model in a different way. Instead of focusing on the single variables and imagining an approximate field of mean value surrounding it, we could have directly focused on approximating the Boltzmann joint distribution itself. This is what we do now by considering the so called *mean field variational approach*.

Idea: let's look for tractable approximations for intractable joint distributions $P(\mathbf{s})$. How? We consider a distribution $Q(\mathbf{s})$ which belongs to a family of tractable distributions and aim at finding one that is as "close" as possible to the exact $P(\mathbf{s})$.
In the Mean Field case, the family is that of fully factorized distributions, i.e. of the type:

$$Q(\mathbf{s}) = \prod_i Q_i(s_i) \quad . \tag{19}$$

For the binary case $s_i \in \pm1$, it turns out that the most general form is:

$$Q_i(s_i) = \frac{1 + s_i m_i}{2} \tag{20}$$

where m_i is not an arbitrary parameter but it is indeed $m_i = \mathbb{E}_Q\left[s_i\right]$.
Exercise: prove it.

Question: how do we measure the distance between the distributions Q and P?
A common metric in probabilistic modeling is the *Kullback-Leibler divergence*:

$$KL(Q||P) := \sum_{\mathbf{s}} Q(\mathbf{s}) \ln \frac{Q(\mathbf{s})}{P(\mathbf{s})} = \mathbb{E}_Q \left[\ln \frac{Q}{P} \right] \quad . \tag{21}$$

Obs1: in general $KL(Q||P) \geq 0$.
Exercise: prove it.

Here we specialize to the class of Boltzmann distributions $P(\mathbf{s}) = \frac{\exp[-\beta H(\mathbf{s})]}{Z}$, where $\mathbf{s}$ is a vector of N binary random variables, $H(\mathbf{s})$ is the energy of the Curie-Weiss system as in Eq. (10), and $Z = \sum_{\mathbf{s}} e^{-H(\mathbf{s})}$ is the partition function.

With this functional form we get:

$$KL(Q||P) = \log Z - \beta F[Q] \tag{22}$$

where

$$\begin{aligned} F[Q] &= E[Q] - \frac{1}{\beta} S[Q] && \textit{variational or Gibbs free energy of Q} && (23) \\ S[Q] &= -\sum_{\mathbf{s}} Q(\mathbf{s}) \log Q(\mathbf{s}) = -\mathbb{E}_Q [\log Q] && \text{entropy of Q} && (24) \\ E[Q] &= \sum_{\mathbf{s}} Q(\mathbf{s}) H(\mathbf{s}) = \mathbb{E}_Q [H] && \text{internal energy of Q} && (25) \\ & && && (26) \end{aligned}$$

Obs2: given that $KL(Q||P) \geq 0$, then $\beta F[Q]$ is a lower bound for $\log Z$ and is an upper bound for $-\log Z =: \beta F_{true}$, where F_{true} is often called Helmholtz free energy[2].
Obs3: notice that the dependence on the distribution $P(\mathbf{s})$ inside $F[Q]$ is contained in the internal energy term $E[Q]$, as the $H(\mathbf{s})$ is the energy of the Boltzmann measure of $P(\mathbf{s})$.

The goal is to minimize the distance between Q and P with respect to the parameters of the distribution Q. Given that $\log Z$ does not contain them, we could then find them by minimizing:

$$Q^*(\mathbf{s}) = \arg\min_{Q \in \mathcal{M}} \{F[Q]\} \tag{27}$$

where the minimum is over the family $\mathcal{M}$ of factorized distributions for the Mean Field case. In other words, we want to find the set of parameters m_i such that the variational free energy of Q is minimal.

For the factorized distributions as in Eq. (20) we have:

$$\begin{aligned} S[Q] &= -\left[\frac{1+m}{2} \log\left(\frac{1+m}{2}\right) + \frac{1-m}{2} \log\left(\frac{1-m}{2}\right) \right] && (28) \\ E[Q] &= -\frac{J}{2} m^2 - hm && (29) \end{aligned}$$

so that the free energy becomes:

$$F[Q] = -\frac{J}{2} m^2 - hm + \frac{1}{\beta} \left[\frac{1+m}{2} \log\left(\frac{1+m}{2}\right) + \frac{1-m}{2} \log\left(\frac{1-m}{2}\right) \right] \tag{30}$$

[2] The inequality $-\frac{\log Z}{\beta} =: F_{true} \leq F[Q]$ is called Bogoliubov-Feynman-Gibbs inequality.

Calculating the derivative with respect to m_i in those expressions, we obtain:

$$m_i = \tanh\left(\frac{J}{2}\sum_{j} m_j + \theta_i\right) \quad . \tag{31}$$

This is the same expression obtained with intuition in Eq. (17).

The previously intractable task to compute exact averages over P has been replaced by solving this set of nonlinear equations, which can often be done in polynomial time.

Exercise: repeat this calculations for a more general Hamiltonian:

$$H(\mathbf{s}) = -\sum_{i<j} J_{ij} s_i s_j - \sum_i s_i h_i \quad . \tag{32}$$

What self-consistent equation do you get?

2.1 Mean Field optimal parameters for Boltzmann distributions

Denoting (as before) $\mathbf{s}_{\setminus i}$ the vector formed by deleting i-th component from $\mathbf{s}$, the Gibbs free energy of the Mean Field distribution (19) can be rewritten as:

$$\begin{aligned} F[Q] &= \mathbb{E}_Q\left[H(\mathbf{s})\right] + \frac{1}{\beta}\sum_{j,s_j\in\pm1} Q_j(x_j)\log Q_j(s_j) && (33)\\ &= \frac{1}{\beta}\sum_{s_i\in\pm1} Q_i(s_i)\log Q_i(s_i) + \frac{1}{\beta}\sum_{j\neq i,s_j\in\pm1} Q_j(s_j)\log Q_j(s_j) + && (34)\\ &\quad + \sum_{s_i\in\pm1} Q_i(s_i)\underbrace{\left[\sum_{\mathbf{s}_{\setminus i}\in\{\pm1\}^{N-1}}\prod_{j\neq i} Q_j(s_j)H(\mathbf{s})\right]}_{=:\mathbb{E}_{\mathbf{s}_{\setminus i}}[H(\mathbf{s})]} \quad . && (35) \end{aligned}$$

Obs1: $\mathbb{E}_{\mathbf{s}_{\setminus i}}[H(\mathbf{s})]$[3] is the conditional expectation of $H(\mathbf{s})$ when we fix s_i, so it is indeed a function of s_i.

Define:

$$P_{\setminus i}(s_i) := \frac{e^{-\beta\,\mathbb{E}_{\mathbf{s}_{\setminus i}}[H(\mathbf{s})]}}{Z_{\setminus i}} \quad , \tag{36}$$

where $\quad Z_{\setminus i} = \sum_{s_i=\pm1} e^{-\beta\,\mathbb{E}_{\mathbf{s}_{\setminus i}}[H(\mathbf{s})]}$.

Question: why is this useful?
Because if we keep isolating terms containing s_i we observe something interesting:

$$\begin{aligned} F[Q] &= \frac{1}{\beta}\sum_{j\neq i,s_j\in\pm1} Q_j(s_j)\log Q_j(s_j) + \sum_{s_i\in\pm1} Q_i(s_i)\left[\mathbb{E}_{\mathbf{s}_{\setminus i}}[H(\mathbf{s})] + \frac{1}{\beta}\log Q_i(s_i)\right]\\ &= \frac{1}{\beta}\sum_{j\neq i,s_j\in\pm1} Q_j(s_j)\log Q_j(s_j) + \frac{1}{\beta}\sum_{s_i\in\pm1} Q_i(s_i)\left[-\log\left(Z_{\setminus i}\frac{\exp\left(-\beta\mathbb{E}_{\mathbf{s}_{\setminus i}}[H(\mathbf{s})]\right)}{Z_{\setminus i}}\right) + \log Q_i(s_i)\right]\\ &= \underbrace{\frac{1}{\beta}\sum_{j\neq i,s_j\in\pm1} Q_j(s_j)\log Q_j(s_j) - \frac{1}{\beta}\log Z_{\setminus i}}_{\text{does not depend on } s_i} + \frac{1}{\beta}KL\left(Q_i\,||\,P_{\setminus i}\right) \quad . \end{aligned} \tag{37}$$

[3] $\sum_{\mathbf{s}_{\setminus i}\in\{\pm1\}^{N-1}}\prod_{j\neq i} Q_j(s_j)H(\mathbf{s}) := \sum_{s_2\in\{\pm1\}}\cdots\sum_{s_N\in\{\pm1\}} Q_2(s_2)\dots Q_N(s_N)H(\mathbf{s})$, where for simplicity we fixed $s_i \equiv s_1$.

This means that minimizing the Gibbs free energy of the factorized distribution Q, is equivalent to minimize the KL divergence between each of the $Q_i(s_i)$ and the $P_{\backslash i}(s_i)$. In other words, the best individual elements of the factorized distribution (19) are the $P_{\backslash i}(s_i)$ as defined in Eq. (36). These are Boltzmann distributions that have as energy function the conditional expected energy given by:

$$\mathbb{E}_{\mathbf{s}_{\backslash i}}[H(\mathbf{s})] := \sum_{s_j \in \pm 1} \prod_{j \neq i} Q_j(s_j) H(\mathbf{s}) \tag{38}$$

Example. Consider the Hamiltonian:

$$H(\mathbf{s}) = -\sum_{i<j} J_{ij} s_i s_j \tag{39}$$

and a factorized distribution Q as in Eq. (20). We can calculate:

$$\begin{aligned}
\mathbb{E}_{\mathbf{s}_{\backslash i}}[H(\mathbf{s})] &= -\sum_{s_j \in \pm 1} \prod_{j \neq i} Q_j(s_j) \left[\sum_{l<j, j, l \neq i} J_{lj} s_l s_j + s_i \sum_{j>i} J_{ij} s_j \right] && (40) \\
&= const - \sum_{s_j \in \pm 1} \prod_{j \neq i} Q_j(s_j) \left[s_i \sum_{j>i} J_{ij} s_j \right] && (41) \\
&= const - s_i \sum_{s_j \in \pm 1} \sum_{j \neq i} \left[\frac{1 + s_j m_j}{2} \right] J_{ij} s_j && (42) \\
&= const - s_i \sum_{j \neq i} J_{ij} \left[\frac{1+m_j}{2} - \frac{1-m_j}{2} \right] && (43) \\
&= const - s_i \sum_{j \neq i} J_{ij} m_j \quad . && (44)
\end{aligned}$$

We can now calculate the partition function:

$$Z_{\backslash i} = \sum_{s_i = \pm 1} e^{\beta s_i \sum_{j \neq i} J_{ij} m_j} = 2 \cosh\left(\beta \sum_{j \neq i} J_{ij} m_j \right) \quad , \tag{45}$$

so that we finally obtain:

$$P_{\backslash i}(s_i) = \frac{1 + \tanh\left(\beta \sum_{j \neq i} J_{ij} m_j\right)}{2} \delta(s_i - 1) + \frac{1 - \tanh\left(\beta \sum_{j \neq i} J_{ij} m_j\right)}{2} \delta(s_i + 1) \tag{46}$$

With this result we can derive an iterative equation to derive the parameters' m_i. In fact, given that $m_i = \mathbb{E}_{P_{\backslash i}}[s_i]$:

$$m_i^{t+1} = \tanh\left(\beta \sum_j J_{ij} m_j^t \right) \quad \Rightarrow \quad \mathbf{m}^{t+1} = \tanh\left(\beta \mathbf{J} \mathbf{m}^t \right) \tag{47}$$

where we have replaced the summation over $j \neq i$ with a summation over all terms, because for large N the extra term does not impact to the leading order; $\mathbf{J}$ is the $N \times N$ coupling matrix.

2.2 Mean Field: summary

- What is the interpretation of the Mean Field approach? We are replacing the fluctuating "field" due to the couplings $h_i = \sum_j J_{ij} s_j$ by an approximation to its mean value.
- This method becomes exact when N, the total number of random variables, is large enough to keep the fluctuations small, assuming that the s_j are weakly dependent through $J_{ij} = \frac{J}{N}$ positive and equal, as in the Curie-Weiss model.
- For general J_{ij} random variables this approximation fails to become exact.