The Many Faces of Mean Field Theory

Anoop Praturu

2026-02-04

There is only one class of problems that physicists can reliably solve: problems where there are no interactions. Even notable cases of interacting problems that *have* been solved, such as 2 gravitating bodies or the 2D Ising model, are solved by transforming the problem into one where there are no interactions (the 2 body problem is reduced to a 1 body problem in an effective potential, and the Ising model is transformed into a system of free fermions). This is troubling because things in the physical world tend to interact with each other[1]. How can we make progress?

Mean field theory (MFT) has been the most reliable way to tackle interacting problems for nearly the last 120 years. The method gets its name from its original formulation where interacting variables, says spins on a lattice in the Ising model, are replaced by their mean values. This makes the problem tractable, since spins never interact directly but instead via the "mean field". However this also imposes a constraint: the mean values that the model *predicts* must agree with the mean value that we put into the interaction. We will see that this "consistency condition" endows mean field theory with a variational structure that allows the effects of interactions to propagate through our solution and yield good approximations.

Despite the simplicity and physical clarity of this formulation, MFT comes in many flavors, where it is often not so clear what the "mean field" actually is. The vast proliferation of mean field methods stems from the many challenging problems MFT has been applied to. Field theory, stat mech, neuroscience, and ML (to name just a few) have each spun out versions of MFT aimed at addressing the specific questions and challenges of that field. The essence of mean field theory that unifies these different forms is to replace an interacting problem with a noninteracting one, and supplement it with a variational principle that injects the physics of interest into our simplified model. The goal of this post is to review the many faces of mean field theory, using the Ising model as a vehicle to introduce them. I hope to tie these different methods together in a logical manner and justify why they all share the name "mean field theory", while at the same time emphasizing the difference in ability and physical content of these models.

I begin with a review of the Ising model, but assume the reader is familiar with the basics of statistical physics. I then introduce the "standard" Weiss Curie mean field theory to set the stage, before showing how this method implies a variational approach to the problem. From there I discuss how mean field methods can be cast in a field theoretic language, and end by discussing improvements to mean field theory that are especially important in neural network style models.

1 The Ising Model

The Ising model is a simple classical spin model for understanding ferromagnetism in solids. We consider a D dimensional hyper-cubic lattice with spins $s_i \pm 1$ on lattice vertices indexed by i. We suppose there are interactions only between neighboring spins in the presence of a possibly spatially varying external magnetic field h_i, and denote the total number of spins by N which we take $\to \infty$ in the thermodynamic limit. The Hamiltonian is then given by:

$$H[s] = -\frac{1}{2}\sum_{i,j} J_{ij} s_i s_j - \sum_i h_i s_i$$

[1] Or so I'm told.

where $J_{ij} = J$ if i and j are neighboring sites, and 0 otherwise[2]. We see that the energy is minimized when all spins take on the same value and collectively magnetize into an ordered state, with J controlling the strength of this cooperative interaction. This cooperation is tempered by thermal fluctuations which seek to destroy order. The competition between these 2 effects makes this a prototypical model for studying phase transitions. The Ising model actually works much better at modeling ferromagnetism in real solids than it has any business to, for deep reasons that we will not cover here (Goldenfeld 2018). At temperature T ($\beta \equiv 1/k_B T$), the object of statistical mechanics is to compute the partition function from this Hamiltonian:

$$Z = \mathrm{Tr}_{\{s\}} e^{-\beta H[s]} = \sum_{s_1=\pm 1} ... \sum_{s_N=\pm 1} e^{-\beta H[s]}$$

since the associated free energy $F \equiv -\beta^{-1} \ln Z$ acts as the generating functional for observables. For example we can compute the average spin[3] as

$$m_i \equiv \langle s_i \rangle = \frac{1}{Z} \mathrm{Tr}_{\{s\}} \left[s_i e^{-\beta H[s]} \right] = \frac{1}{Z} \mathrm{Tr}_{\{s\}} \left[\beta^{-1} \frac{\partial}{\partial h_i} e^{-\beta H[s]} \right] = \beta^{-1} \frac{1}{Z} \frac{\partial Z}{\partial h_i} = -\frac{\partial F}{\partial h_i}$$

The derivatives of F encode the thermodynamics of the system by generating thermal averages. The magnetization is often referred to as an *order parameter*, since it becomes non-zero in the ordered state. A quantity of fundamental importance is the connected correlation function

$$G_{ij} \equiv \left\langle s_i s_j \right\rangle_c = \left\langle s_i s_j \right\rangle - \left\langle s_i \right\rangle \left\langle s_j \right\rangle = \left\langle (s_i - \langle s_i \rangle)(s_j - \left\langle s_j \right\rangle) \right\rangle \tag{1}$$

Correlations are inherently driven by interactions. If the magnetization fluctuates at a site i it will "polarize" its neighbors to fluctuate in the same direction, and this effect will propagate through the system. This is the first sign that the correlation function may be difficult to handle in mean field theory, which seeks to eschew interactions. G_{ij} arises from the free energy as (dropping $\{s\}$ subscripts for clarity)

$$G_{ij} = \frac{1}{Z} \mathrm{Tr} \left[s_i s_j e^{-\beta H} \right] - \frac{1}{Z^2} \mathrm{Tr} \left[s_i e^{-\beta H} \right] \mathrm{Tr} \left[s_j e^{-\beta H} \right] = \frac{\beta^{-2}}{Z} \frac{\partial^2 Z}{\partial h_i \partial h_j} - \frac{\beta^{-2}}{Z^2} \frac{\partial Z}{\partial h_i} \frac{\partial Z}{\partial h_j}$$

$$\beta^{-2} \left[\frac{1}{Z} \frac{\partial^2 Z}{\partial h_i \partial h_j} + \frac{\partial}{\partial h_j} \left(\frac{1}{Z} \right) \frac{\partial Z}{\partial h_i} \right] = \beta^{-2} \frac{\partial^2 \ln Z}{\partial h_j \partial h_i} = -\beta^{-1} \frac{\partial^2 F}{\partial h_j \partial h_i}$$

A closely related concept is the local susceptibility

$$\chi_{ij} \equiv \frac{\partial m_i}{\partial h_j} = -\frac{\partial^2 F}{\partial h_i \partial h_j} = \beta G_{ij}$$

This relationship between G_{ij} and χ_{ij} is referred to variously as the "fluctuation dissipation theorem" or the "static susceptibility sum rule". Though G and χ are mathematically equivalent, their physical interpretations are quite different. G measures how the *internal* thermal fluctuations propagate through the system, while χ measures how the system's response to an *external* perturbation propagates. It's intuitive that these quantities should be related, but we'll see that the difference in their physical content will rear its head in mean field theory.

In 1 dimension, the Ising model is easily solved via the Transfer Matrix method or high temperature expansion (Bellac and Barton 1992). In 2 dimensions, the Ising model was first exactly solved by Lars Onsager in a *tour de force* of theoretical physics (Onsager 1944). Alternative solutions (SCHULTZ, MATTIS, and LIEB 1964; Kac and Ward 1952) followed shortly, though all solutions are in the presence of 0 external field. No exact solution has been found in 3 dimensions. No solutions are needed in $D \geq 4$ dimensions because the results of MFT are exact in those dimensions (Goldenfeld 2018) (a fact that we will not touch on here). Our focus will not be on these exact solutions, but on the zoo of approximate methods of solution that originated almost 40 years before Onsager's solution, and have seen continous development into the present day.

[2]We adopt this notation so that we will be able to express future calculations in matrix form rather than clunky nearest neighbor sums. This notation is also amenable to generalizing to different models where J_{ij} is fully connected or randomly distributed as in a spin glass.

[3]I will often refer to m_i as the *local magnetization* or sometimes just the *magnetization*, though the latter sometimes refers to the *bulk magnetization* of the system obtained by averaging m_i over space. I hope it will be clear from context which I am referring to.

1.1 Other Kinds of Ising Models

Before we move on to *solving* the Ising model I should briefly mention the plethora of other spin based models. The most famous extension of the Ising model is the *Spin Glass*, which is an Ising model on a lattice in which the couplings are randomly distributed $J_{ij} \sim \mathcal{N}(0, J_0^2)$ if i and j are nearest neighbors. This model gives rise to an immensely complex spin glass phase due to the fact that it allows for *geometric frustration*: combinations of positive and negative bonds can lead to configurations in which one cannot simultaneously put all spin pairs into their lowest energy states. This leads to a proliferation of metastable states in the low temperature phase. Randomness in the couplings is typically referred to as *quenched disorder*, meaning that the randomness is "frozen in" while the microscopic spin variables fluctuate on a much faster thermal timescales. Mathematically we express this by computing the free energy at fixed couplings, and then averaging over the couplings afterwards. This is accomplished via the "replica trick" (Mézard, Parisi, and Virasoro 1987) by using the following identity

$$\langle \ln Z \rangle = \lim_{n \to 0} = \frac{\langle Z^n \rangle - 1}{n}$$

The spin glass is often studied in terms of its infinite range counterpart, the Sherrington-Kirkpatrick model[4]. Specifically we have $J_{ij} \sim \mathcal{N}(0, J_0^2/N)$ for *all* pairs i, j, and the $1/N$ scaling in the variance is to ensure that the energy scales extensively as N. The statistical physics of the spin glass phase and the replica methods used to compute it forms a fascinating subject that we will not dive into here, but I mention it because of the deep connection between spin glasses and neural neworks. It was shown that certain models of emergent collective memory minimize an energy function that is identical to the Sherrington Kirkpatrick Hamiltonian (Hopfield 1982). The spin states act as binary codes, and the network utilizes the many metastable states of the spin glass to store these codes. In this post I strive to leave hamiltonians in their most general form with J_{ij} unspecified, so that the results we derive can be applied broadly across statistical physics style lattice models, disordered spin systems, and neural networks. I also emphasize that these MFT techniques extend far beyond the Ising model, we only restrict ourselves here so that we have a unified vehicle with which to introduce methods.

2 Weiss-Curie Mean Field Theory

We'll start by looking at the original formulation of MFT by Curie and Weiss, in which the "mean field" nature of the solution is most clear. A spin can be decomposed into its static mean value plus a dynamic fluctuation: $s_i = m_i + \delta s_i$. We can then write the interacting term as

$$s_i s_j = (m_i + \delta s_i)(m_j + \delta s_j) = m_i s_j + m_j s_i - m_i m_j + \delta s_i \delta s_j$$

The first 2 terms show how the spins couple to the mean field, the third is a constant which can be ignored, and the final term shows how the fluctuations couple to each other. The idea behind mean field theory is to drop the interaction between fluctuations, effectively taking $s_i s_j \to s_i m_j + s_j m_i$. The Hamiltonian is then

$$H_{MF} = -\sum_{i,j} J_{ij} s_i m_j - \sum_i h_i s_i \equiv -\sum_i h_i^e s_i \tag{2}$$

Where we define the effective external field as:

$$h_i^e \equiv h_i + \sum_j J_{ij} m_j$$

The field that a spin feels at site i is the existing external field plus the sum of the magnetizations of all of its neighbors. The mean field hamiltonian is now completely decoupled and the partition function sum factorizes

$$Z = \prod_{i=1}^{N} \left(\sum_{s_i = \pm 1} e^{\beta h_i^e s_i} \right) = 2^N \prod_{i=1}^{N} \cosh(\beta h_i^e) \implies F = -\beta^{-1} \sum_{i=1}^{N} \ln\left(2 \cosh(\beta h_i^e)\right)$$

[4]The infinite range model is in fact the mean field theory for the spin glass on a finite dimensional lattice. We will explain this fact in the final section.

As promised, the problem was made tractable by removing interactions. We must now ensure that the MFT is self consistent:

$$m_i = \frac{\sum_{s_i=\pm 1} s_i e^{\beta s_i h_i^e}}{2\cosh(\beta h_i^e)} = \tanh(\beta h_i^e)$$

which gives the following self consistency constraint equation:

$$m_i = \tanh\left(\beta\left(h_i + \sum_j J_{ij} m_j\right)\right) \tag{3}$$

Setting the external field uniformly to 0 we can take all the $m_i = m$ to be equal by symmetry, and denoting the lattice coordination number by $z = 2D$ we have

$$m = \tanh(\beta z J m) \tag{4}$$

which can be solved graphically.

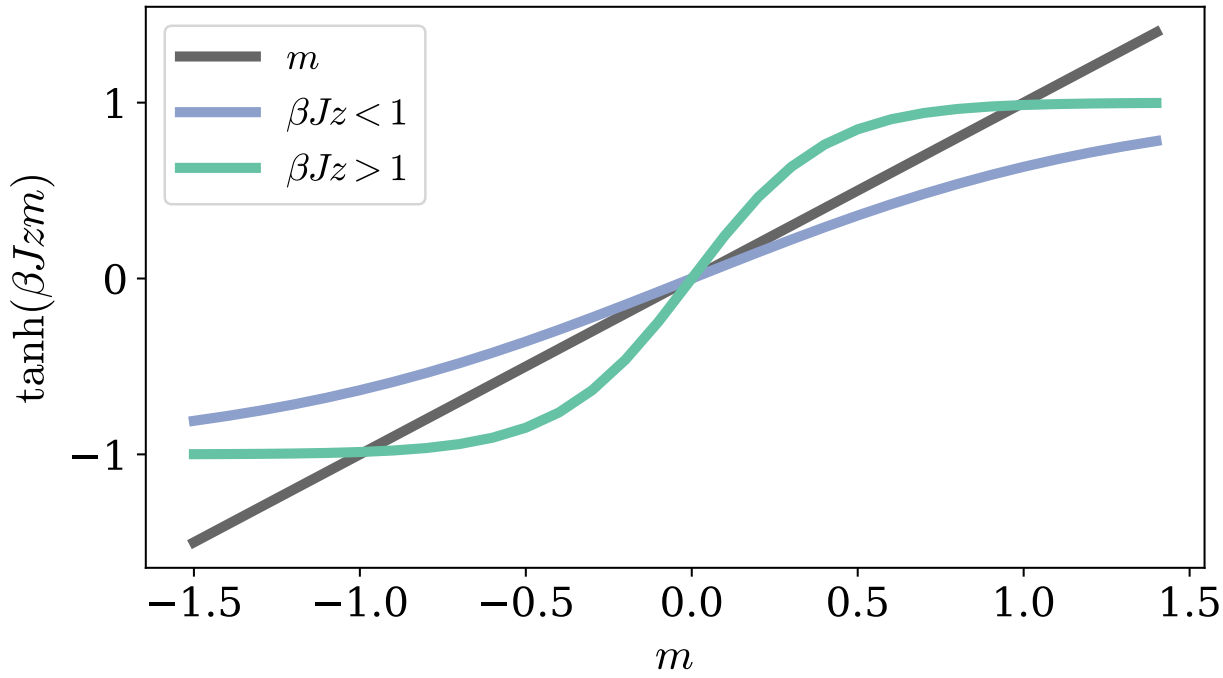

Figure 1: Equation 4 can be solved graphically by plotting the left and right hand sides of the equation and looking for intersections. There are only non-zero solutions for m when the slope of the tanh is greater than 1.

When the slope of the tanh function at the origin is less than 1 there is only the trivial solution at $m = 0$. When the slope is greater than 1, there exist solutions at non-zero m, so we expect a phase transition at $\beta_c = (2JD)^{-1}$ where all of the spins collectively magnetize in the same direction and the system *orders*. Whether the system chooses the positive or negative solution for m is the subject of the fascinating topic of spontaneous symmetry breaking (Goldenfeld 2018).

2.1 Correlation Functions and Susceptibility

We have shown that MFT predicts a phase transition from a disordered (0 net magnetization) to ordered (collectively magnetized) phase at low temperatures. What do the correlation functions and susceptibility look like in these phases? Notice that from Equation 1 we can write the correlation function as $G_{ij} = \langle \delta s_i \delta s_j \rangle$,

but in our mean field theory we explicitly set $\delta s_i \delta s_j = 0$ so it seems $G_{ij} = 0$. This makes sense: a theory which explicitly ignores fluctuations cannot compute the thermodynamics of fluctuations. The correlation function is however non-zero onsite ($i = j$) since $s_i = \pm 1$, so $s_i^2 = 1$, and $\langle s_i^2 \rangle = 1$. Thus we must have

$$G_{ij} = \delta_{ij}(1 - m_i^2) \tag{5}$$

Should the susceptibility then also only be non-zero onsite? Let's proceed by differentiating Equation 3 directly.

$$\chi_{ij} = \frac{\partial m_i}{\partial h_j} = \beta(1 - m_i^2)\left(\delta_{ij} + \sum_k J_{ik}\chi_{kj}\right) \tag{6}$$

We can formally solve this equation in matrix form (I provide an explicit solution via fourier transforms in the Appendix Section 7.1) by defining the diagonal matrix $\chi^0_{ii} = \beta(1 - m_i^2)$. Then

$$= (\mathbf{I} - {}^0\mathbf{J})^{-1}{}^0 = {}^0 + {}^0\mathbf{J}^0 + {}^0\mathbf{J}^0\mathbf{J}^0 + ... \tag{7}$$

^{0} alone agrees with Equation 5, but higher order terms in the expansion generate additional contributions. The first term gives non-zero values for nearest neighbor elements of , and the next gives next-next-nearest neighbor contributions, and so on. It seems that correlations and susceptibility are no longer equivalent in mean field theory. Physically, our model cannot make predictions about internal thermal fluctuations since we explicitly ignore them, but Equation 3 has injected enough information about interactions that the model can predict how external perturbations propagate. Mathematically, the distinction is more subtle. Something that is obscured by Weiss-Curie mean field theory is that the role of free energy, and the manner in which we calculate thermal averages of order parameters changes drastically in mean field theory. We will now present an alternative approach to MFT that brings these features to light.

3 Variational Mean Field Theory

The tractability of MFT comes from the fact that the mean field Hamiltonian Equation 2 does not have interactions between neighboring spins. We saw however that the form of H_{MF} is not arbitrary and h_i^e is constrained by the consistency conditions. One question we could ask is: can we find a non-interacting theory that does better than mean field theory? Suppose we introduced a trial hamiltonian

$$H_g = -\sum_i g_i s_i$$

parameterized by g_i. Could we find a set of values for the g_i that gave better predictions than mean field theory? Let $p(s) = Z^{-1}\exp(-\beta H)$ be the true probability distribution over microstates from the interacting Ising Hamiltonian, and let $q(s) = Z_g^{-1}\exp(-\beta H_g)$ be the distribution from the trial Hamiltonian. We can find the best possible trial Hamiltonian by minimizing the KL divergence between the 2 distributions.

$$\mathrm{KL}(q||p) = \mathrm{Tr}_{\{s\}}\left(q(s)\ln\frac{q(s)}{p(s)}\right) = \beta F_g - \beta F + \beta\left\langle H - H_g\right\rangle_q$$

Where $\langle\cdot\rangle_q$ denotes averages with respect to the trial distribution $q(s)$. Using the fact that the KL divergence ≥ 0 and $F_g = \left\langle H_g\right\rangle_q - TS_q$ (where S_q denotes the entropy of q) we obtain the Bogoliubiov inequality

$$F \leq F_g + \beta\left\langle H - H_g\right\rangle_q = \langle H\rangle_q - TS_q \equiv \mathcal{F}[g]$$

We define the *variational free energy* as $\mathcal{F}$, and see from the above that it acts as an upper bound on the true free energy. We can interpret minimizing the KL divergence as closing the gap between $\mathcal{F}$ and the true free energy F by minimizing $\mathcal{F}$. Computing the averages gives

$$\mathcal{F}[g] = -\beta^{-1}\sum_i \ln(2\cosh(\beta g_i)) - \frac{1}{2}\sum_{ij} J_{ij}m_i m_j - \sum_i h_i m_i + \sum_i g_i m_i \tag{8}$$

where $m_i = \tanh(\beta g_i)$. differentiating gives

$$\frac{\partial \mathcal{F}}{\partial g_i} = -\beta(1-m_i^2)\sum_j J_{ij}m_j - \beta(1-m_i^2)(h_i - g_i)$$

We seek the minimum by setting to 0 which gives

$$g_i^* = h_i + \sum_j J_{ij}m_j \tag{9}$$

or multiplying by β and taking the tanh

$$m_i = \tanh\left(\beta\left(h_i + \sum_j J_{ij}m_j\right)\right) \tag{10}$$

The solution of the variational problem is exactly the mean field solution, and g_i^* has the exact same form as h_i^e. Mean field theory is the *best possible* non-interacting approximation.

3.1 What is Free Energy in Mean Field Theory?

We define the mean field free energy as

$$F_{MF}[h] = \min_g \mathcal{F}[g;h] = \mathcal{F}[g^*;h]$$

Note that this is *not* the factorized free energy F_g; that was only used as a vehicle to compute averages within a tractable model. F_{MF} is the "correct" approximate free energy in the sense that it is as close as possible to the true free energy, even though it is not calculated directly from a partition function sum over microscopic variables. We can see the difference explicitly by plugging Equation 9 into Equation 8

$$F_{MF}[h] = F_{g^*} - \frac{1}{2}\sum_{ij} J_{ij}m_i m_j$$

F_g only includes interactions through g^*, whereas F_{MF} explicitly accounts for them. We recover

$$-\frac{\partial F_{MF}}{\partial h_i} = -\sum_j \left.\frac{\partial \mathcal{F}(g;h)}{\partial g_j}\right|_{g_j=g_j^*} \left.\frac{\partial g_j}{\partial h_i}\right|_{g_j=g_j^*} - \frac{\partial \mathcal{F}(g^*;h)}{\partial h_i} = -\frac{\partial \mathcal{F}(g^*;h)}{\partial h_i} = m_i(g^*) \tag{11}$$

since $\mathcal{F}(g^*;h)$ is stationary by definition, but notice that differentiating F_{MF} wasn't actually necessary to generate the thermal averages of the s_i. We were able to calculate m_i in Equation 10 as a consequence of minimizing $\mathcal{F}$. This highlights a major conceptual shift in mean field theory:

> Mean field theory calculates thermodynamics by solving a variational problem. The free energy does *not* play the role of a generating function, but instead organizes our approximation by acting as a target for our variational problem.

In the same way that minimizing action give you classical equations of motion, minimizing free energy gives you an equation of state. This perspective explains the apparent contradiction from the end of the last section: the susceptibility and correlation do not agree because the generating function arguments used to relate them break down in this framework. Something else that arises operationally is that **mean field theory replaces integration with optimization**.

3.2 Gibbs Free Energy

If optimization is what gives us order parameters m_i, can we formulate MFT as an optimization directly over m_i instead of over the trial external fields g_i? To do this, we must Legendre transform from a fixed field ensemble to a fixed magnetization ensemble:

$$G_{MF}[m] = F_{MF}[h(m)] + \sum_i h_i(m)m_i \tag{12}$$

$h(m)$ is computed by inverting Equation 11. The physical interpretation of the second term is as the external field necessary to enforce that $\langle s_i \rangle = m_i$, since the m_i are specified as independent variables in the ensemble, rather than derived. Again leveraging Equation 11 we easily see that

$$\frac{\partial G_{MF}}{\partial m_i} = \sum_j \frac{\partial F_{MF}}{\partial h_j}\frac{\partial h_j}{\partial m_i} + \sum_j \frac{\partial h_j}{\partial m_i} m_j + h_i = h_i$$

Using the fact that under the trial distribution $P(s_i = 1) = \frac{1}{2}(1 + m_i)$ we can write an explicit expression for G_{MF}:

$$G_{MF}[m_i] = \beta^{-1} \sum_i \left[\frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} \right] - \frac{1}{2} \sum_{ij} J_{ij} m_i m_j$$

And minimizing with respect to m_i recovers Equation 10. The formulation in terms of G_{MF} also sheds light on the susceptibility:

$$\chi_{ij}^{-1} = \frac{\partial h_i}{\partial m_j} = \frac{\partial^2 G_{MF}}{\partial m_i \partial m_j}$$

The susceptibility (specifically its inverse) is encoded in the *curvature* of the Gibbs free energy landscape. This is because the Hessian matrix on the right hand side measures the curvature of a surface at it's fixed point. We saw that mean field theory neglects fluctuations; this is intimately related to it's variational nature. By replacing integration with optimization our model is only informed by the free energy at a single point, rather than the whole landscape. Fluctuations are what allow the ensemble to explore this whole space. We see hints of this now in the susceptibility: in order to compute χ_{ij} we need to know the *shape* of the minimum, not just it's location. MFT cannot compute the correlation function because it does not internally have access to this broader geometric structure, but it *can* probe this second order structure via the response function. This also hints at a possible way to extend mean field theory. If we could find a way to systematically introduce fluctuations in a controlled perturbative manner, could we improve our model and possibly compute the correlation function?

4 Field Theoretic Approach

The difficulty with handling fluctuations in a spin model is that they are discrete. There is no controlled way to introduce interactions "a little bit at a time"; they are all or nothing. If we want to use perturbative methods, we need a model with continuous degrees of freedom that can be expanded infinitesimally. Remarkably, discrete spin variables can be turned into continuous variables via the *Hubbard-Stratanovich transform*:

$$\exp\left[\frac{1}{2}\sum_{ij} J_{ij} s_i s_j\right] = \frac{1}{(2\pi)^{N/2}\sqrt{\det(J)}} \int_{-\infty}^{\infty} \prod_i d\phi_i e^{-\frac{1}{2}\sum_{ij}\phi_i (J^{-1})_{ij}\phi_j + \sum_i \phi_i s_i}$$

This follows from completing the square in the integrand and performing the Gaussian integral, but in the reverse direction it has the effect of decoupling the spins from each other. Spins interact indirectly via an intermediate field ϕ with Gaussian self interactions. The partition function then has the form:

$$Z \propto \int \prod_i d\phi_i e^{-\frac{1}{2\beta}\sum_{ij}\phi_i J_{ij}^{-1}\phi_j} \mathrm{Tr}\left[e^{\sum_i s_i(\phi_i + \beta h_i)}\right]$$

In this form the spin contribution factorizes and the trace can be performed to give

$$Z = \propto \int \prod_i d\phi_i e^{-S[\phi]}$$

Where the action $S[\phi]$ is given by

$$S[\phi] = \frac{1}{2\beta}\sum_{ij}\phi_i J_{ij}^{-1}\phi_j - \sum_i \ln(2\cosh(\phi_i + \beta h_i)) \tag{13}$$

Notice that this action scales approximately as $\sim N$, so by the saddle point approximation we expect that the integral will be dominated by its value at the minimum of the action. We can then organize our approximation by finding the minimum, and then expanding fluctuations about this point in increasing order. To find the minimum, denoted by ϕ_i^*, we demand that

$$\frac{\partial S}{\partial \phi_i} = \frac{1}{\beta}\sum_j J_{ij}^{-1}\phi_j^* - \tanh(\phi_i^* + \beta h_i) = 0$$

defining $m_i = \beta^{-1}\sum_j J_{ij}^{-1}\phi_j^*$ at the minimum we get

$$m_i = \tanh\left(\beta\left(h_i + \sum_j J_{ij}\phi_j\right)\right)$$

These are exactly the consistency equations of mean field theory! It makes sense physically that ϕ_i plays the role of the magnetization m_i: it is the field that mediates interactions between spins. We again see the mean field equations arise as the solution to a variational problem, but it is of a different nature this time. The action in Equation 13 has no guarantee of being an upper bound on the true free energy. It is minimized because that is where we expect the dominant contribution to the free energy to come from, but it is not the "best" approximate free energy in the same sense as it was in variational MFT. We can now incorporate the lowest order interactions between fluctuations by expanding to second order about the minimum $\phi_i = \phi_i^* + \delta\phi_i$:

$$S[\delta\phi] \approx S[\phi^*] + \frac{1}{2}\sum_{ij}\delta\phi_i\Gamma_{ij}\delta\phi_j + O(\delta\phi^3)$$

where

$$\Gamma_{ij} = \frac{1}{\beta}J_{ij}^{-1} - (1 - m_i^2)\delta_{ij}$$

We could do the integral and solve for the free energy, but we recognize that the second order action defines a Gaussian distribution so we can immediately read off the correlation function as

$$\langle \delta\phi_i \delta\phi_j \rangle = \Gamma_{ij}^{-1} \equiv (\Gamma_0 - \Pi)_{ij}^{-1}$$

where we have defined the "bare" kernel $\Gamma_0 \equiv \beta^{-1}J^{-1}$ and polarization $\Pi \equiv \text{diag}(1 - m_i^2)$. Notice that the polarization is what we derived for the on-site correlation function in Equation 5. It represents how a spin responds to its *own* fluctuations. The bare kernel on the other hand is a new contribution to the correlation function that arises directly from the coupling of spins. Defining the correlation matrix $G_{ij} \equiv \langle \delta\phi_i \delta\phi_j \rangle$ we can expand the matrix inversion to give

$$G = \Gamma_0^{-1} + \Gamma_0^{-1}\Pi\Gamma_0^{-1} + \Gamma_0^{-1}\Pi\Gamma_0^{-1}\Pi\Gamma_0^{-1} + ...$$

Up to matrix multiplications by J^{-1} to convert between ϕ_i and m_i, this is exactly the expansion we obtained for the susceptibility in Equation 7! The ability of the Gaussian field theory to give a non-trivial answer for the correlation function depends crucially on the fact that we were able to include interactions between fluctuations, but still in a tractable manner. I emphasize once more the subtle difference in the variational nature of the Gaussian theory from the previous formulation of mean field theory. The Gaussian theory doesnt replace integration with optimization, it instead uses optimization to find the largest contribution to the free energy for which integration is tractable. The approximate free energy we obtain still acts as a generating function, and does not necessarily bound the true free energy.

Note that while we have kept the calculation above in its most general form by leaving J as a matrix, the Ising model on a lattice can be solved explicitly in Fourier space by exploiting the translationally invariant nature of the problem. This is essentially identical to the susceptibility calculation we do in Section 7.1. A more "statistical field theory" flavored approach is to take the continuum limit $\phi_i \to \phi(x)$ and express Z as a functional integral over fields. The key insight there is to show that the J matrix turns into a gradient of the field. While these are interesting approaches from a field theoretic perspective, the general form is much more relevant to modern systems of interest such as neural networks and spin glasses.

4.1 The Random Phase Approximation

What is the interpretation of the series expansion for the correlation function? Though they are equivalent, we work with the susceptibility χ_{ij} for the magnetization since it is more physical. Recall

$$= (\mathbf{I} - {}^0\mathbf{J})^{-1} {}^0 = {}^0 + {}^0\mathbf{J}^0 + {}^0\mathbf{J}^0\mathbf{J}^0 + ...$$

Where $\beta^{-1}\chi^0 = \text{diag}(1-m_i^2)$. χ^0 describes how a single spin responds to an external perturbation in isolation, without any effects from other spins. Per Equation 3 this change in m_i will induce a change in its neighbors, which will then feedback into m_i and re-perturb it. This is what the second term in the expansion encodes. Higher terms continue this chain reaction: the n^{th} order term describes all possible length n chains of influence that start and end on site i. The susceptibility (or correlation function) at a single site requires us to sum over all possible paths in the interaction network that could affect site i. This is referred to in the physics literature as the "Random Phase Approximation"

In the language of Feynman diagrams, we can think of χ^0 as a local irreducible bubble diagram connected to a site, and J as a propagator that connects different sites. The expansion is then a sum of "bubble-chain" diagrams: local bubble insertions to describe how the field responds, and chain of propagators to describe how this response moves through the system. Although the resummed series can be solved exactly (see Appendix Section 7.1), this interpretation sheds light on the structure of our approximation. Mean field theory only retains interaction structure with a specific linear chain topology. Branching or crossing paths of influence are entirely possible, but they couple interactions in a way that we discard in mean field theory.

5 The TAP equations

All of our different approaches to mean field theory have, in one form or another, produced the same equation of state Equation 3. The physical content of these models is that the "effective field" felt by a spin at site i is given by

$$h_i^e = h_i + \sum_{ij} J_{ij} m_j$$

Notice that the m_j determining the field at i depend on m_i itself. h_i^e is supposed to account for the influence of the environment on spin i, but it has included information about spin i. The distinciton between spin at i and environment has not been made precise. Physically, the spin at i polarizes it neighbors which then feed back into i and produce an "overcounting" effect. In the standard Ising model on a square lattice this is negligible; it is a second order effect summed over a small number of local neighbors. In "dense" networks which are highly connected this can have a non-negligible effect since the contribution is summed over a macroscopic number of spins. The TAP equations are the next highest order correction to Equation 3 that accounts for this effect. They first arose in the context of the infinite range spin glass model, where one can show that they are the only correction necessary. The equations, and higher order corrections, can be derived systematically via the *Plefka expansion* of the Gibbs free energy, but I'll present a more physical approach based on the argument given above.

5.1 The Cavity Method

I present an intuitive sketch of the derivation here. A plethora of more rigorous approaches can be found in (Opper and Saad 2001). In order to determine the effective field at i self consistently without feedback, we should consider the system with the spin at i removed. i.e. with a *cavity* at i. Define

$$m_j^{\setminus i} \equiv \langle s_j \rangle_{\setminus i}$$

to be the magnetization at j in the system where the spin at i is removed. Formally, you can think of this as an Ising model where $J_{ij} = 0 \ \forall j$ at fixed i. Then we should have

$$h_i^{e\setminus i} = h_i + \sum_j J_{ij} m_j^{\setminus j}$$

We can complete the system by expressing $m_j^{\setminus i}$ in terms of m_j and writing the equation of state as $m_i = \tanh(\beta h_i^{e\setminus i})$. To do this, imagine we add the spin back to site i, then to first order the effective field felt at j changes by $\delta h_j^e = J_{ij} s_i \approx J_{ij} m_i$. We can calculate the change this induces in $m_j^{\setminus i}$ via the susceptibility

$$m_j - m_j^{\setminus i} \approx \sum_k \chi_{jk}^{\setminus i} \delta h_k^e$$

notice that $\chi^{\setminus i}$ above is already multiplied by a perturbative term, so to the order we are working in we can consider only the lowest order contributions to χ:

$$\sum_k \chi_{jk}^{\setminus i} \delta h_k^e \approx \chi_{jj}^{\setminus i} \delta h_j^e = \beta(1 - (m_j^{\setminus i})^2)\delta h_j^e \approx \beta(1 - m_j^2)\delta h_j^e$$

We only retain the local contribution χ_{jj}, since longer chains of interaction only contribute at higher order, and take $m_j^{\setminus i} \approx m_j$ since their difference is $O(\delta h_j^e)$. Combining these gives

$$m_j^{\setminus j} \approx m_j - \beta m_i J_{ij}(1 - m_j^2) + O((\delta h^e)^2)$$

Using the fact that J_{ij} is symmetric we can plug this back in to the equation of state to give

$$m_i = \tanh\left(\beta\left(h_i + \sum_j J_{ij} m_j - \beta m_i \sum_j J_{ij}^2 (1 - m_j^2)\right)\right) \tag{14}$$

The second term is referred to as the *Onsager reaction term* and describes the correction to the effective field from removing the feedback overcounting. Lets think about the structure of the reaction term in the language of RPA that we developed earlier. The reaction is given by

$$m_i \sum_j J_{ij}^2 \chi_{jj}^0$$

In the language of diagrams this looks like: propagate from $i \to j$ via J_{ij}, local bubble interaction χ^0 at j, propagate back $j \to i$ via J_{ji}. The TAP equations modify the equation of state by considering how the local effects of polarization feed back into the system via the shortest possible chains.

6 When Does Mean Field Theory Work?

Throughout our winding tour of the many faces of mean field theory we have taken great pains to understand exactly what we are retaining and what we are throwing away in our approximations. One question we have not addressed is: when are we justified in using mean field theory? Under what circumstances are our approximations appropriate? Let's close out our discussion of MFT by tracing out the boundary of its regime of applicability.

- **Dense Networks**: Consider an "infinite range" Ising model in which $J_{ij} = J/N$ for every possible pair of spins. Writing the Hamiltonian as $H = -\frac{J}{2N}(\sum_i s_i)^2 - h\sum_i s_i$, we can transform into a continuum theory with a single variable given by

  $$Z \propto \int d\phi e^{-NS(\phi)}, \quad S(\phi) = \frac{J}{2}\phi^2 - \ln(2\cosh(\beta(h + J\phi)))$$

  As $N \to \infty$ we can use the saddle point method to solve the integral by evaluating the integrand at the minimum action. So the free energy is given *exactly* by the action at it's minimum, and fluctuations don't contribute. Intuitively, an individual spin in a fully connected network recieves signals from every other spin. These overlapping signals tend to "average out" and produce a mean field interaction. If a fluctuation emerges it is immediately communicated to the rest of the network. In a sparsely connected network, by contrast, local signals can vary widely, and fluctuations have to propagate via

local interactions which can allow distant parts of the network to differ strongly. At first glance, a system in which every variable interacts with every other variable may seem daunting in its complexity. Mean field theory offers a different perspective: highly connected interaction networks can sometimes be *more* tractable if their structure encourages the mean field conditions. Another implication is that one can construct a "mean field" version of a problem by finding it's infinite range analog. For example, the canonical mean field approach to studying spin glasses is the Sherrington Kirkpatrick model: an infinite range Ising model with randomly distributed couplings.

- **Small Fluctuations**: A point we have returned to throughout, in many different forms, is that the main approximation of mean field theory is to treat fluctuations as small. In the decomposition $s_i = m_i + \delta s_i$, we roughly want our theory to satisfy something along the lines of $\langle \delta s_i^2 \rangle \ll m_i^2$. In statistical physics, this is made quantitative by the *Ginzburg Criterion*:

$$E_G \equiv \frac{|\int_V d^d r G(r)|}{\int_V d^d r m(r)^2}$$

  Where G and m are computed according to the field theoretic formulation of mean field theory. To find regimes where MFT is not valid, we look for places where E_G is large or possibly divergent. Remarkably, MFT can predict it's own demise! In the regions where MFT itself predicts large fluctuations, we know we cannot trust the theory. In the statistical physics of the Ising model we find that near the critical point E_G diverges because the correlation length diverges, but only for $D < 4$ (this gives rise to the notion of upper critical dimension). In those cases MFT can predict the physics of the bulk phases well, but necessarily fails to model the phase transition correctly.
- **Things are Gaussian**: We saw in Section 4 that mean field theory is equivalent to finding the best approximate gaussian distribution to fit our model. One way to think about the applicability of mean field theory is to ask how Gaussian your system is. This is common in the theory of neural networks: in extremely wide networks at initialization the input to a given neuron is the sum of a bunch of indepent random activations, and so approaches gaussianity by the central limit theorem. The mean field approximation then amounts to replacing all distributions with tractable gaussians whose covariances have been fit to match the actual correlations of the network (Schoenholz et al. 2017). Using more sophisticated RG techniques (that were originally designed to make up for the shortcomings of mean field theory) one can asses quantitatively from the data how important non-gaussian terms are, and whether or not they can be ignored (Bradde and Bialek 2017).

Despite the masterclass in physics that led to the development of renormalization, mean field theory still stands as the dominant technique to understand interacting problems. Part of this is because the conditions under which mean field theory holds are surprisingly broad and relevant. Another reason is that it is often the *only* means by which we can make progress. But really, I attribute the success of mean field theory to its many faces. MFT is not one single idea or algorithm. It is a set of overlapping, sometimes contradictory, yet constantly evolving ideas and techniques[5]. Every problem admits its own MFT, that requires its own set of assumptions, approximations, and iterative improvements. While the techniques described in this post are nice, I hope what you take away is something deeper and more structural.

7 Appendix

7.1 Exact Mean Field Theory Susceptibility Calculation

Let us consider Equation 6 in the case where the external field is uniform so that $m_i = m$. Then we can write

$$\sum_k \left(\frac{\delta_{ik}}{\beta(1-m^2)} - J_{ik} \right) \chi_{kj} = \delta_{ij} \tag{15}$$

Suppose the lattice has spacing a, volume $V = Na^D$ and let $\vec{r}_i$ denote the spatial position of the lattice site i, so for example $\chi_{ij} = \chi(|\vec{r}_i - \vec{r}_j|)$. Since the system is translationally invariant all functions on the lattice only

[5]This post has not come close to touching on all of the different mean field theories that have been developed. Some notable ones I have enjoyed learning about are dynamical mean field theory, the replica methods in spin glasses, and a deeper exploration of the TAP equations and their interpretation in terms of message passing.

depend on the separation between lattice points, and the matrices are diagonalized in fourier space. The lattice Fourier transform is given by

$$\tilde{f}(\vec{q}) = \sum_{\vec{r}} f(\vec{r}) e^{-i\vec{q}\cdot\vec{r}}, \quad f(\vec{r}) = \frac{1}{V}\sum_{\vec{q}} \tilde{f}(\vec{q}) e^{i\vec{q}\cdot\vec{r}_j}$$

It follows that we can write the Kronecker Delta in fourier space as

$$\delta_{ij} = \frac{1}{V}\sum_{\vec{q}} e^{i\vec{q}\cdot(\vec{r}_i - \vec{r}_j)}$$

To take the fourier transform of J_{ij} we note that $J(\vec{r}) = J$ iff $\vec{r}$ is a lattice vector of length a. Thus for each dimension d the fourier transform gets a contribution $Je^{\pm iq_d a}$, where the $\pm$ comes from the forward and backward direction that a lattice vector can point along a given dimension. This gives

$$\tilde{J}(\vec{q}) = 2J\sum_{d=1}^{D} \cos(q_d a)$$

Consider then the following matrix product

$$\sum_k J_{ik}\chi_{kj} = \frac{1}{V^2}\sum_{\vec{r}_k}\sum_{\vec{q},\vec{q'}} \tilde{\chi}(\vec{q})\tilde{J}(\vec{q'}) e^{i(\vec{q'}\cdot\vec{r}_i - \vec{q}\cdot\vec{r}_j)} e^{i\vec{r}_k\cdot(\vec{q}-\vec{q'})} = \frac{1}{V}\sum_{\vec{q}} \tilde{\chi}(\vec{q})\tilde{J}(\vec{q}) e^{i\vec{q}\cdot(\vec{r}_i - \vec{r}_j)}$$

Where we have used the fact that

$$\frac{1}{V}\sum_{\vec{r}_k} e^{i\vec{r}_k\cdot(\vec{q}-\vec{q'})} = \delta(\vec{q} - \vec{q'})$$

to perform the sums over $\vec{r}_k$ and $\vec{q'}$. Similarly we have

$$\sum_k \frac{\delta_{ik}}{\beta(1-m^2)}\chi_{kj} = \frac{1}{V}\sum_{\vec{q}} \frac{1}{\beta(1-m^2)}\tilde{\chi}(\vec{q}) e^{i\vec{q}\cdot(\vec{r}_i - \vec{r}_j)}$$

equating the fourier transforms on the left and right hand sides of Equation 15 then gives

$$\tilde{\chi}(\vec{q}) = \frac{1}{\chi_0^{-1} - 2J\sum_{d=1}^{D}\cos(q_d a)}$$

where $\chi_0 \equiv \beta(1-m^2)$. Consider the behavior on scales much larger than the lattice spacing where $q_d a \ll 1$. Then to lowest order in q we have

$$\tilde{\chi}(\vec{q}) \approx \frac{1}{\chi_0^{-1} - 2JD + Ja^2|\vec{q}|^2}$$

Define $\kappa^2 \equiv (\chi_0^{-1} - 2JD)/(Ja^2)$, $c \equiv Ja^2$, and the correlation length $\xi \equiv 1/\kappa$. Then

$$\chi(\vec{r}) = \frac{1}{c}\int \frac{d^D\vec{q}}{(2\pi)^D}\frac{e^{i\vec{q}\cdot\vec{r}}}{|\vec{q}|^2 + \kappa^2} = \frac{1}{Ja^2}\frac{\kappa^{D/2-1}}{(2\pi)^{D/2}}\frac{K_{D/2-1}(\kappa r)}{r^{D/2-1}}$$

where $r = |\vec{r}|$ and K_ν is the modified bessel function of the second kind. When $r \gg \xi$ we can take the asymptotic form of the Bessel function to get

$$\chi(\vec{r}) \propto \frac{e^{-r/\xi}}{r^{(D-1)/2}}$$

susceptibility (and correlation) looks power law out to the correlation length ξ after which correlations get exponentially suppressed. You can think about ξ as the average "size" of a fluctuation. Consider near the critical point where $m \approx 0$, and let $T_c \equiv 2JD$ be the mean field transition temperature we derived in Section 2. Then

$$\xi^2 = \frac{(Ja)^2}{T - T_c}$$

We see that the correlation length diverges as $|T - T_c|^{-1/2}$ at the critical point. At the phase transition correlations decay purely as a power law and there are fluctuations at all scales. This behavior is a signal the mean field theory breaks down at the critical point, and renormalization methods are required. The scale free behavior also points to the structure of the RG solution, but that is a story for another time.

Bellac, Michel Le, and Gabriel Barton. 1992. *Quantum and Statistical Field Theory*. Oxford University Press.

Bradde, Serena, and William Bialek. 2017. "PCA Meets RG." *Journal of Statistical Physics* 167 (3–4): 462–75. https://doi.org/10.1007/s10955-017-1770-6.

Goldenfeld, Nigel. 2018. *Lectures on Phase Transitions and the Renormalization Group*. CRC Press.

Hopfield, John J. 1982. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities." *Proceedings of the National Academy of Sciences* 79 (8): 2554–58.

Kac, M., and J. C. Ward. 1952. "A Combinatorial Solution of the Two-Dimensional Ising Model." *Phys. Rev.* 88 (December): 1332–37. https://doi.org/10.1103/PhysRev.88.1332.

Mézard, Marc, Giorgio Parisi, and Miguel Angel Virasoro. 1987. *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*. Vol. 9. World Scientific Publishing Company.

Onsager, Lars. 1944. "Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition." *Phys. Rev.* 65 (February): 117–49. https://doi.org/10.1103/PhysRev.65.117.

Opper, Manfred, and David Saad. 2001. *Advanced Mean Field Methods: Theory and Practice*. MIT press.

Schoenholz, Samuel S., Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. 2017. "Deep Information Propagation." https://arxiv.org/abs/1611.01232.

SCHULTZ, T. D., D. C. MATTIS, and E. H. LIEB. 1964. "Two-Dimensional Ising Model as a Soluble Problem of Many Fermions." *Rev. Mod. Phys.* 36 (July): 856–71. https://doi.org/10.1103/RevModPhys.36.856.