From Naive Mean Field Theory to the TAP Equations

Manfred Opper and Ole Winther

The MIT Press
Cambridge, Massachusetts
London, England

We give a basic introduction to three different MF approaches which will be discussed on a more advanced level in other chapters of this book. We discuss the Variational, the Field Theoretic and the TAP approaches and their applications to a Boltzmann machine type of Ising model.

1 Introduction

Mean field (MF) methods provide tractable approximations for the computation of high dimensional sums and integrals in probabilistic models. By neglecting certain dependencies between random variables, a closed set of equations for the expected values of these variables is derived which often can be solved in a time that only grows polynomially in the number of variables. The method has its origin in Statistical Physics where the thermal fluctuations of particles are governed by high dimensional probability distributions.

In the field of probabilistic modeling, the MF approximation is often identified as a special kind of the *variational approach* in which the true intractable distribution is approximated by an optimal factorized one. On the other hand, a variety of other approximations with a "mean field" flavor are known in the Statistical Physics community. However, compared to the variational approach the derivation of these other techniques seem to be less "clean". For instance, the "field theoretic" MF approaches may lack a clearcut probabilistic interpretation because of the occurrence of auxiliary variables, integrated in the complex plane. Hence, one is often unable to turn such a method into an exact bound. Nevertheless, as the different contributions to this book show, the power of non-variational MF techniques should not be ignored.

This chapter does not aim at presenting any new results but rather tries to give a basic and brief introduction to three different MF approaches which will be discussed on a more advanced level in other chapters of this book. These are the *Variational*, the *Field Theoretic* and the *TAP* approaches. Throughout the chapter, we will explain the application of these methods for the case of an Ising model (also known as a Boltzmann machine in the field of Neural Computation).

Our review of MF techniques is far from being exhaustive and we expect that other methods may play an important role in the future. Readers who want to learn more about Statistical Physics techniques

and the MF method may consult existing textbooks e.g. (16; 19; 33). A more thorough explanation of the variational method and its applications will be given in the chapters (5; 7; 9) of this book. A somewhat complementary review of advanced MF techniques is presented in the next chapter (32).

2 The Variational Mean Field Method

Perhaps the best known derivation of mean field equations outside the Statistical Physics community is the one given by the *Variational Method*. This method approximates an intractable distribution $P(\mathbf{S})$ of a vector $\mathbf{S} = (S_1, \ldots, S_N)$ of random variables by $Q(\mathbf{S})$ which belongs to a family $\mathcal{M}$ of tractable distributions. The distribution Q is chosen such that it minimizes a certain distance measure $D(Q, P)$ within the family $\mathcal{M}$.

To enable tractable computations, $D(Q, P)$ is chosen as the *relative entropy*, or *Kullback-Leibler divergence*

$$KL(Q||P) = \sum_{\mathbf{S}} Q(\mathbf{S}) \ln \frac{Q(\mathbf{S})}{P(\mathbf{S})} = \left\langle \ln \frac{Q}{P} \right\rangle_Q , \tag{1}$$

where the bracket $\langle \ldots \rangle_Q$ denotes an expectation with respect to Q. Since $KL(Q||P)$ is not symmetric in P and Q, one might wonder if $KL(P||Q)$ would be a better choice (this question is discussed in the two chapters of (28; 1)). The main reason for choosing (1) is the fact that it requires only computations of expectations with respect to the tractable distribution Q instead of the intractable P.

We will specialize on the class of distribution P that are given by

$$P(\mathbf{S}) = \frac{e^{-H[\mathbf{S}]}}{Z} , \tag{2}$$

where $\mathbf{S} = (S_1, \ldots, S_N)$ is a vector of binary (spin) variables $S_i \in \{-1, +1\}$ and

$$H[\mathbf{S}] = -\sum_{i<j} S_i J_{ij} S_j - \sum_i S_i \theta_i . \tag{3}$$

Finally, the normalizing partition function is

$$Z = \sum_{\mathbf{S}} e^{-H[\mathbf{S}]} . \tag{4}$$

We are interested both in approximations to expectations like $\langle S_i \rangle$ as well as in approximations to the value of the *free energy* $-\ln Z$.

Inserting P into (1), we get

$$KL(Q||P) = \ln Z + E[Q] - S[Q] \tag{5}$$

where

$$S[Q] = -\sum_{\mathbf{S}} Q(\mathbf{S}) \ln Q(\mathbf{S}) \tag{6}$$

is the entropy of the distribution Q (not to be confused with the random variable S) and

$$E[Q] = \sum_{\mathbf{S}} Q(\mathbf{S}) H[\mathbf{S}] \tag{7}$$

is called the *variational energy*.

The mean field approximation is obtained by taking the approximating family $\mathcal{M}$ to be all product distributions, i.e.

$$Q(\mathbf{S}) = \prod_j Q_j(S_j) \ . \tag{8}$$

For $S_i \in \{-1, +1\}$, the most general form of the Q_j's is obviously of the form:

$$Q_j(S_j; m_j) = \prod_j \frac{(1 + S_j m_j)}{2} \tag{9}$$

where the m_j's are variational parameters which are identified as the expectations $m_j = \langle S_j \rangle_Q$.

Using the statistical independence of the S_j's with respect to Q, the variational entropy is found to be

$$S[Q] = -\sum_i \left\{ \frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} \right\} \ . \tag{10}$$

and the variational energy reduces to

$$E[Q] = \langle H[\mathbf{S}] \rangle_Q = -\sum_{i<j} J_{ij} m_i m_j - \sum_i m_i \theta_i \ . \tag{11}$$

Although the partition function Z cannot be computed efficiently, it will not be needed because it does not depend on Q. Hence, all we have to

do is to minimize the *variational free energy*

$$F[Q] = E[Q] - S[Q] \; . \tag{12}$$

Differentiating (12) with respect to the m_i's gives the set of N *Mean Field Equations*

$$m_i = \tanh\left(\sum_j J_{ij} m_j + \theta_i\right) \; , \qquad i = 1, \ldots, N \; . \tag{13}$$

The intractable task of computing exact averages over P has been replaced by the problem of solving the set (13) of nonlinear equations, which often be done in a time that grows only polynomially with N. Note, that there might be many solutions to (13) and some of them may not even be local minima of (12) but rather saddles. Hence, solutions must be compared by their value of the variational free energy $F[Q]$.

As an extra bonus of the variational MF approximation we get an upper bound on the exact free energy $-\ln Z$. Since $KL(Q||P) \geq 0$, we have from (5)

$$-\ln Z \leq E[Q] - S[Q] = F[Q] \; . \tag{14}$$

Obviously, the mean field approximation takes into account the couplings J_{ij} between the random variables but neglects statistical correlations, in the sense that $\langle S_i S_j \rangle_Q = \langle S_i \rangle_Q \langle S_j \rangle_Q$. To get some more intuition about the effect of this approximation, we can compare the mean field equations for $m_i = \langle S_i \rangle_Q$ (13) with a set of *exact* equations which hold for the *true* distribution P (2). It is not hard to prove the so-called *Callen equations* (see e.g. chapter 3 of (19))

$$\langle S_i \rangle = \left\langle \tanh\left(\sum_j J_{ij} S_j + \theta_i\right) \right\rangle , \qquad i = 1, \ldots, N \; . \tag{15}$$

Unfortunately both sides of (15) are formulated in terms of expectations (we have omitted the subscript) with respect to the difficult P. While in (15) the expectation is *outside* the nonlinear tanh function, the approximate (13) has the expectation *inside* the tanh. Hence, the MF approximation replaces the fluctuating "field" $h_i = \sum_j J_{ij} S_j$ by (an approximation) to its *mean field*. Hence, estimating the variance of h_i may give us an idea of how good the approximation is. We will come

back to this question later.

3 The Linear Response Correction

Although the product distribution $Q(\mathbf{S})$ neglects correlations between the random variables, there is a simple way of computing a non-vanishing approximation to the covariances $\langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle$ based on the MF approach. By differentiating

$$\langle S_i \rangle = Z^{-1} \sum_{\mathbf{S}} S_i \, e^{-H[\mathbf{S}]} \tag{16}$$

with respect to θ_j, we obtain the *linear response* relation

$$\frac{\partial \langle S_i \rangle}{\partial \theta_j} = \langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle \; . \tag{17}$$

(17) holds only for expectations with respect to the true P but not for the approximating Q. Hoping that the MF method gives us a reasonable approximation for $\langle S_i \rangle$, we can compute the MF approximation to the left hand side of (17) and get a nontrivial approximation to the right hand side. This approximation has been applied to Boltzmann machines learning (11) and independent component analysis (8).

4 The Field Theoretic Approach

Another way of obtaining a mean field theory is motivated by the idea that we often have better approximations to the performance of integrals than to the calculation of discrete sums. If we can replace the expectations over the random variables S_i by integrations over auxiliary "field variables", we can approximate the integrals using the Laplace or saddle-point methods.

As an example, we consider a simple Gaussian transformation of (2). To avoid complex representations we assume that the matrix $\mathbf{J}$ is positive definite so that we can write

$$\exp\left[\frac{1}{2}\sum_{ij} S_i J_{ij} S_j\right] =$$

$$\frac{1}{(2\pi)^{N/2}\sqrt{\det(J)}} \int_{-\infty}^{\infty} \prod_i dx_i \; e^{-\frac{1}{2}\sum_{ij} x_i (\mathbf{J}^{-1})_{ij} x_j + \sum_i x_i S_i} \; . \tag{18}$$

This transformation is most easily applied to the partition function Z (4) yielding

$$Z \propto \int \prod_i dx_i \; e^{-\frac{1}{2}\sum_{ij} x_i (\mathbf{J}^{-1})_{ij} x_j} \prod_j \left\{ \sum_{S_j} e^{S_j (x_j + \theta_j)} \right\} \; , \tag{19}$$

where we have omitted some constants. In this representation, the sums over binary variables factorize and can be carried out immediately with the result

$$Z \propto \int \prod_i dx_i \; e^{\Phi(\mathbf{x})} \; , \tag{20}$$

where

$$\Phi(\mathbf{x}) = -\frac{1}{2} \sum_{ij} x_i (\mathbf{J}^{-1})_{ij} x_j + \sum_j \ln \left(2 \cosh(x_j + \theta_j) \right) \; . \tag{21}$$

Hence, we have transformed a high-dimensional sum into a high dimensional non-Gaussian integral. Hoping, that the major contribution to the integral comes from values of the function Φ close to its maximum, we replace the integral (20) by

$$Z \approx e^{\Phi(\mathbf{x}^0)}, \tag{22}$$

where $\mathbf{x}^0 = \arg\max \Phi(\mathbf{x})$. This is termed the Laplace approximation. Setting the gradient $\nabla_{\mathbf{x}} \Phi(\mathbf{x})$ equal to zero, we get the set of equations

$$\sum_j (\mathbf{J}^{-1})_{ij} x_j^0 = \tanh(x_i^0 + \theta_i) \; . \tag{23}$$

A comparison of (23) with (13) shows that by identifying the auxiliary variables x_i^0 with the mean fields via

$$x_i^0 \equiv \sum_j J_{ij} m_j \; , \tag{24}$$

we recover the same mean field equations as before. This is easily understood from the fact that we have replaced the integration variables x_i by constant values. This leaves us with a partition function for the

same type of factorizing distribution

$$Q(\mathbf{S}) \propto \prod_j e^{S_j(x_j^0+\theta_j)} \tag{25}$$

(written in a slightly different form) that we have used in the variational approach.

Hence, it seems we have not gained anything new. One might even argue that we have lost something in this derivation, the bound on the free energy $-\ln Z$. It is not clear how this could be proved easily within the Laplace approximation. However, we would like to argue that when interactions between random variables are more complicated than in the simple quadratic model (7), the field-theoretic approach decouples the original sums in a very simple and elegant way for which there may not be an equivalent expression in the variational method. This can often be achieved by using a Dirac δ-function representation which is given by

$$1 = \int dh\ \delta(h-x) = \int \frac{dhd\hat{h}}{2\pi}\ e^{i\hat{h}(h-x)}\ , \tag{26}$$

where the $i = \sqrt{-1}$ in the exponent should not be confused with a variable index. The transformation can be applied to partition functions of the type

$$\begin{aligned} Z &= \sum_{\mathbf{S}} \prod_j f\left(\sum_k J_{jk} S_k\right) \\ &= \sum_{\mathbf{S}} \int \prod_j \left\{ dh_j f(h_j)\ \delta\left(h_j - \sum_k J_{jk} S_k\right)\right\} \qquad (27) \\ &= \int \prod_j \left(\frac{dh_j d\hat{h}_j}{2\pi}\right) e^{-i\sum_j \hat{h}_j h_j} \prod_k \left\{ \sum_{S_k} e^{iS_k \sum_j J_{jk}\hat{h}_j} \right\}\ . \qquad (28) \end{aligned}$$

Since the functions in (28) are no longer positive (in fact, not even real), the search for a maximum in Φ must be replaced by the Saddle-point method where (after a deformation of the path of integration in the the complex plane), one looks for values of h and $\hat{h}$ for which the corresponding exponent is stationary.

In general, the field theoretic MF approach does not have an equivalent variational formulation (in fact, depending on the way the auxiliary fields are chosen, we may get different MF formulations). Hence, it is

unclear if the approximation to Z will lead to a bound for the free energy. While there is no general answer so far, an example given in one of the chapters of this book (22) indicates that in some cases this may still be true.

A further important feature of the saddle-point approximation is the fact that it can be systematically improved by expanding Φ around the stationary value. The inclusion of the quadratic terms may already give a dramatic improvement. Applications of these ideas to graphical models can be found in this book (22; 2).

5 When does MFT become exact?

We have seen from the Callen equation (15) that the simple MF approximation neglects the fluctuations of the fields

$$h_i = \sum_j J_{ij} S_j \ , \tag{29}$$

which are sums of random variables. In the interesting case where N, the total number of variables S_j is large one might hope that fluctuations could be small assuming that the S_j are weakly dependent. We will compute crude estimates of these fluctuations for two extreme cases.

• Case I:
All couplings J_{ij} are positive and equal. In order to keep the fields h_i of order $\mathcal{O}(1)$ when N grows large, we set $J_{ij} = J_0/N$. This model is known as the mean field ferromagnet in Statistical Physics. If we make the crude approximation that all variables S_j are independent, the variances $\mathrm{Var}(J_{ij}S_j) = J_0^2 \left(1 - \langle S_j \rangle^2 \right) / N^2$ of the individual terms in (29) simply add to a total variance of the fields $\mathrm{Var}(h_i) = \mathcal{O}(1/N)$ for $N \to \infty$. Hence, in this case the MF approximation becomes exact. A more rigorous justification of this result can be obtained within the field theoretic framework of the previous section. The necessary Gaussian transformation for this case is simpler than (18) and reads

$$\exp\left[\frac{J_0}{2} N \sum_{ij} S_i S_j\right] = \exp\left[\frac{J_0}{2N} (\sum_i S_i)^2\right] \propto \int dx \, e^{-\frac{N}{2J_0} x^2} \, e^{x \sum_i S_i} \ . \tag{30}$$

Inserting (30) into the partition function (4) shows that Laplace's

method for performing the single integral over x is justified for $N \to \infty$ by the occurrence of the factor N in the exponent.
In practical applications of MF methods, the couplings J_{ij} are usually related to some observed data and will not be constant but may rather show a strong variability. Hence, it is interesting to study the

- Case II:

The J_{ij}'s are assumed to be independent random variables (for $i < j$) with zero mean. Setting $\theta_i = 0$ for simplicity, we are now adding up N terms in (29) which have roughly equal fractions of positive and negative signs. To keep the h_i's of order 1, the magnitude of the J_{ij}'s should then scale like $1/\sqrt{N}$. With the same arguments as before, neglecting the dependencies of the S_j's, we find that the variance of h_i is now $\mathcal{O}(1)$ for $N \to \infty$ and the simple MF approximation fails to become exact.

As will be shown in the next section, the failure of the "naive" mean field theory (13) in case II can be cured by a adding a suitable correction. This leads us to the TAP mean field theory which is still a closed set of equations for the expectations $\langle S_i \rangle$. Under some conditions on the variance of the J_{ij}'s it is believed that these mean field equations are exact for Case II in the limit $N \to \infty$ with probability 1 with respect to a random drawing of the J_{ij}'s.

In fact, it should be possible to construct an exact mean field theory for any model where the J_{ij}'s are of "infinite range". The phrase *infinite range* is best understood if we assume for a moment that the spins S_i are located at sites i on a finite dimensional lattice. If the J_{ij}'s do not decay to zero when the distance $||i - j||$ is large, we speak of an infinite range model. In such cases, the "neighbors" S_j of S_i which contribute dominantly to the field h_i (29) of a spin S_i are not clustered in a small neighborhood of site i but are rather distributed all over the system. In such a case, we can expect that dependencies are weak enough to be treated well in a mean field approximation. Especially, when the connections J_{ij} between two arbitrary spins S_i and S_j are completely random (this includes sparse as well as extensive connectivities), the model is trivially of infinite range.

6 TAP equations I : The cavity approach

The TAP mean field equations are named after D.J. Thouless, P.W. Anderson and R.G. Palmer (29) who derived a MF theory for the Sherrington-Kirkpatrick (SK) model (26). The SK model is of the type (3) where the couplings J_{ij} are independent Gaussian random variables for $i < j$ with variance J_0/N. For simplicity, we set the mean equal to zero. We will give two derivations in this chapter. A further derivation and generalizations is presented in another chapter of this book (10).

Perhaps the most intuitive one is the cavity method introduced by Parisi and Mézard (16). It is closely related to the Bethe approximation (3) which is an exact mean field theory on a tree.

Our goal is to derive an approximation for the marginal distribution $P_i(S_i)$ for each spin variable. We begin with the exact representation

$$P_i(S_i) = \sum_{\mathbf{S}\backslash Si} P(\mathbf{S}) \propto \sum_{\mathbf{S}\backslash Si} e^{S_i(\sum_j J_{ij}S_j + \theta_i)} P(\mathbf{S}\backslash S_i) \; . \tag{31}$$

$P(\mathbf{S}\backslash S_i)$ equals the joint distribution of the $N-1$ spins $\mathbf{S}\backslash S_i$ for an auxiliary system, where S_i has been removed (by setting the J_{ij}'s equal to zero for all $j \neq i$). If the graph of nonzero J_{ij}'s would be a tree, i.e., if it would contain no loops, the S_j's would be fully independent after being disconnected from S_i. In this case, the joint distribution $P(\mathbf{S}\backslash S_i)$ would factorize into a product of individual marginals $P_{j\backslash i}(S_j)$. ¿From this, one would obtain immediately the marginal distribution as

$$P_i(S_i) \propto \prod_j \left[\sum_{S_j} e^{S_i J_{ij} S_j} P_{j\backslash i}(S_j) \right] \quad . \tag{32}$$

Within the tree assumption one could proceed further (in order to close the system of equations) by applying the same procedure to each of the auxiliary marginals $P_{j\backslash i}(S_j)$ and expressing them in terms of their neighbors (excluding S_i). This would lead us directly to the Belief-propagation (BP) algorithm (21) for recursively computing a set of "messages" defined by

$$m_{ji}(S_i) = \sum_{S_j} e^{S_i J_{ij} S_j} P_{j\backslash i}(S_j) \; . \tag{33}$$

This approach as well as its applications will be presented in more detail

in other chapters (4; 30; 25; 32). The route from the BP method to the TAP equations is presented in (13).

We will follow a different route which leads to considerable simplifications by utilizing the fact that the SK model is fully connected. Going back to the formulation (31), we see that the only dependence between S_i and the other variables S_j is through the field $h_i = \sum_j J_{ij} S_j$. Hence, it is possible to rewrite the marginal distribution (32) in terms of the joint distribution of S_i and h_i

$$P(S_i, h_i) \propto e^{S_i(h_i+\theta_i)} \; P(h_i \backslash S_i) \; , \tag{34}$$

where we have introduced the "cavity" [1] distribution of h_i as

$$P(h_i \backslash S_i) = \sum_{\mathbf{S} \backslash S_i} \; \delta(h_i - \sum_j J_{ij} S_j) \; P(\mathbf{S} \backslash S_i) \; . \tag{35}$$

We get

$$P_i(S_i) = \frac{\int dh_i \; e^{S_i(h_i+\theta_i)} \; P(h_i \backslash S_i)}{\sum_{S_i} \int dh_i \; e^{S_i(h_i+\theta_i)} \; P(h_i \backslash S_i)} \; . \tag{36}$$

For the SK model the independence used in (32) does not hold, but one may argue that it can be safely replaced in the following by *sufficiently weak* correlations. In the limit $N \to \infty$, we assume that this is enough to invoke a central limit theorem for the field h_i and replace (35) by the simple Gaussian distribution [2]

$$P(h_i \backslash S_i) \approx \frac{1}{\sqrt{2\pi V_i}} \exp\left(-\frac{(h_i - \langle h_i \rangle_{\backslash i})^2}{2V_i} \right) \; , \tag{37}$$

in the computation of (36). We have denoted an average over the cavity distribution by $\langle \cdot \rangle_{\backslash i}$. Using (37) within (36) we get immediately

$$\langle S_i \rangle = \tanh\left(\langle h_i \rangle_{\backslash i} \right) \; , \qquad i = 1, \ldots, N \; , \tag{38}$$

as the first part of the TAP equations. (38) should be compared to the corresponding set of "naive" MF equations (13) which can be written as

$$\langle S_i \rangle_{\text{naive}} = \tanh\left(\langle h_i \rangle \right) \; , \qquad i = 1, \ldots, N \; . \tag{39}$$

In order to close the system of equations we have to express the cavity

1 the name is is derived from the physical context, where h_i is the magnetic field at the cavity which is left when spin i is removed from the system.

2 The cavity method for a model with finite connectivity is discussed in (15).

expectations $\langle h_i \rangle_{\setminus i}$ and the variances V_i in terms of the full expectations

$$\langle h_i \rangle = \sum_j J_{ij} \langle S_j \rangle \ . \tag{40}$$

Within the Gaussian approximation (37) we get

$$\langle h_i \rangle = \sum_{S_i} \int dh_i \ P(S_i, h_i) \ h_i = \langle h_i \rangle_{\setminus i} + V_i \langle S_i \rangle \ . \tag{41}$$

Hence, only the variances V_i of the cavity field remain to be computed. By definition, they are

$$V_i = \sum_{j,k} J_{ij} J_{ik} \left(\langle S_j S_k \rangle_{\setminus i} - \langle S_j \rangle_{\setminus i} \langle S_k \rangle_{\setminus i} \right) \ . \tag{42}$$

Since the J_{ij}'s are modeled as independent random variables we argue that the fluctuations of the V_i's with respect to the random sampling of the couplings can be neglected for $N \to \infty$ and we can safely replace V_i by

$$\overline{V_i} = \sum_j \overline{J_{ij}^2} \overline{\left(1 - \langle S_j \rangle_{\setminus i}^2 \right)} \approx \frac{J_0}{N} \sum_j \left(1 - \langle S_j \rangle^2 \right) \ , \tag{43}$$

where the bar denotes an average over the distribution of the J_{ij}'s. Note, that by the independence of the couplings, the averages over the J_{ij}'s and the terms $\langle S_j \rangle_{\setminus i}$ factorize. To get the last expression in (43) we have assumed that both the fluctuations the effect of removing S_i can be neglected in the sum. ¿From equations (38),(41) and (43) we get the TAP equations for the SK model

$$\langle S_i \rangle = \tanh \left(\sum_j J_{ij} \langle S_j \rangle - J_0 (1 - q) \langle S_i \rangle + \theta_i \right) \ , \tag{44}$$

where $q = \frac{1}{N} \sum_j (1 - \langle S_j \rangle^2)$. Equations (44) differ from the simple or "naive" MF equations (13) by the correction $-J_0(1-q)\langle S_i \rangle$, which is usually called the *Onsager Reaction Term*. Although the simple MF approximation and the TAP approach are based on weak correlations between random variables, the TAP approach makes this assumption only when computing the distribution of the *cavity* field h_i, i.e., for the case when S_i is disconnected from the system. The Onsager term is the difference between $\langle h_i \rangle$ and the cavity expectation $\langle h_i \rangle_{\setminus i}$ (compare (38)

and (39)) and takes into account the reaction of the neighbors S_j due to the correlations created by the presence of S_i.

A full discussion about why and when (44) yields an exact mean field theory for the SK model is subtle and goes beyond the scope of this chapter. Interested readers are referred to (16). We can only briefly touch the problems. The main property in deriving the TAP equations is the assumption of weak correlations expressed as

$$\lim_{N\to\infty} \frac{1}{N^2} \sum_{k,l} (\langle S_k S_l \rangle - \langle S_k \rangle \langle S_l \rangle)^2 = 0 \; . \tag{45}$$

which can be shown to hold for the SK model when the size of the couplings J_0 is sufficiently small. In this case, there is only a single solution to (44). Things become more complicated with increasing J_0. Analytical calculations show that one enters a complex free energy landscape, i.e. a (spin glass) phase of the model where one has exponentially many (in N) solutions. This corresponds to a multimodal distribution with many equally important modes. (45) is no longer valid for a full average but for local averages within a single mode. Numerical solutions to the TAP equations turn out to be extremely difficult in this region (17) and not all of them can be accepted because they violate the positive definiteness of the covariance matrix $\langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle$. For a setup of the cavity approach in this complex region see chapter V of (16) and in this volume (31) which also discusses its application to stochastic dynamics.

Finally, we want to mention the work of M. Talagrand (see e.g. (27)) who is developing a rigorous mathematical basis for the cavity method.

7 TAP equations II: Plefka's Expansion

Plefka's expansion (23) is a method for deriving the TAP equations by a systematic perturbative computation of a function $G(\mathbf{m})$ which is minimized by the vector of expectations $\mathbf{m} = \langle \mathbf{S} \rangle$.

To define $G(\mathbf{m})$, we go back to the minimization of the variational free energy (12), and do not restrict the distributions Q to be product distributions. We minimize $F(Q) = E[Q] - S[Q]$ in two steps: In the first step, we perform a constrained minimization in the family of all distributions Q_m which satisfy

$$\langle \mathbf{S} \rangle_Q = \mathbf{m} \tag{46}$$

where $\mathbf{m}$ is fixed. We define the *Gibb's Free Energy* as the constrained minimum

$$G(\mathbf{m}) = \min_Q \{E[Q] - S[Q] \mid \langle \mathbf{S} \rangle_Q = \mathbf{m}\} \ . \tag{47}$$

In the second step, we minimize G with respect to the vector $\mathbf{m}$. Since the full minimizer of $F[Q]$ equals the true distribution P, the minimizer of $G(\mathbf{m})$ coincides with the vector of true expectations $\langle S_i \rangle$.

Constrained optimization problems like (47) can be transformed into unconstrained ones by introducing appropriate Lagrange multipliers h_i where we have to minimize

$$E[Q] - S[Q] - \sum_i h_i \left(\langle S_i \rangle_Q - m_i \right) \ , \tag{48}$$

and the h_i's must be chosen such that (46) holds. (48) is again of the form of a variational free energy (12) where $H[Q]$ is replaced by $H[Q] - \sum_i h_i S_i$. Hence, the minimizing distribution is just

$$Q_{\mathbf{m}}(\mathbf{S}) = Z^{-1}(\mathbf{h}) \ e^{-H[\mathbf{S}] + \sum_i h_i S_i} \tag{49}$$

with $Z^{-1}(\mathbf{h}) = \sum_{\mathbf{S}} \ e^{-H[\mathbf{S}] + \sum_i h_i S_i}$. Inserting this solution back into (47) yields

$$G(\mathbf{m}, \mathbf{h}) = \sum_i h_i m_i - \ln \sum_{\mathbf{S}} e^{-H[\mathbf{S}] + \sum_i h_i S_i} \ . \tag{50}$$

The condition (46) on the h_i can be finally introduced by the variation on the vector $\mathbf{h}$

$$G(\mathbf{m}) = \max_{\mathbf{h}} \left\{ \sum_i h_i m_i - \ln \sum_{\mathbf{S}} e^{-H[\mathbf{S}] + \sum_i h_i S_i} \right\} \ . \tag{51}$$

This follows by setting the gradient with respect to $\mathbf{h}$ equal to zero and checking the matrix of second derivatives. The geometric meaning of the function $G(\mathbf{m})$ within Amari's *Information Geometry* is highlighted in the chapters (28; 1).

Why do we bother solving the more complicated 2-stage optimization process, when computing $G(\mathbf{m})$ is as complicated as computing the exact free energy $F[P] = -\ln Z$? It turns out, that a useful perturbation expansion of $G(\mathbf{m})$ with respect to the complicated coupling term $H[\mathbf{S}]$ can be developed. We replace $H[\mathbf{S}]$ by $\lambda H[\mathbf{S}]$ in (51) and expand (setting

$\theta_i = 0$ for simplicity)

$$G(\mathbf{m}) = G_0(\mathbf{m}) + \lambda G_1(\mathbf{m}) + \frac{\lambda^2}{2!} G_2(\mathbf{m}) + \dots \tag{52}$$

with $G_n = \frac{\partial^n}{\partial \lambda^n} G(\mathbf{m})\big|_{\lambda=0}$. The computation of the G_n is a bit tricky because one also has to expand the Lagrange parameters h_i which maximize (51) in powers of λ. However, the first two terms are simple. To zeroth order we obtain $m_i = \tanh(h_i^0)$ and

$$G_0(\mathbf{m}) = \sum_i \left\{ \frac{1+m_i}{2} \ln \frac{1+m_i}{2} + \frac{1-m_i}{2} \ln \frac{1-m_i}{2} \right\} . \tag{53}$$

The calculation of the first order term is also simple, because the first derivative of G at $\lambda = 0$ can be written as an expectation of $H[\mathbf{S}]$ with respect to a factorizing distribution with mean values $\langle S_i \rangle = m_i$. We get

$$G_1(\mathbf{m}) = -\sum_{i<j} J_{ij} m_i m_j . \tag{54}$$

A comparison of the first two terms with (12), (10) and (11) shows that we have already recovered the simple mean field approximation. One can show that the second order term in the expansion is

$$G_2(\mathbf{m}) = -\frac{1}{2} \sum_{ij} J_{ij}^2 (1-m_i)^2 (1-m_j)^2 . \tag{55}$$

Minimizing (52) with respect to $\mathbf{m}$ for $\lambda = 1$ and keeping only terms up to second order, yields the TAP expansion (44)[3].

Plefka's method allows us to recover the TAP equations from a systematic expansion, which in principle allows for improvements by adding higher order terms. Corrections of this type can be found in other chapters in this book (32; 28). Moreover, the approximate computation of $G(\mathbf{m})$ can be used to get an approximation for the free energy $-\ln Z = F[P] = \min_{\mathbf{m}} G(\mathbf{m})$ as well.

For the SK model, Plefka (23) shows that all terms beyond second order in the λ expansion (52) can be neglected with probability 1 (with respect to random drawings of the J_{ij}'s) for $N \to \infty$ as long as we are not in the complex (spin glass) phase of the model.

3 One also has to replace J_{ij}^2 by its average.

8 TAP equations III: Beyond the SK model

The TAP approach is special among the other mean field methods in the sense that one has to make probabilistic assumptions on the couplings J_{ij} in (3) in order to derive the correct MF equations. This causes extra problems because the magnitude of the Onsager correction term will depend on the distribution of J_{ij}'s. E.g., both the SK model and the Hopfield model (6) belong to the same class of models (3) but are defined by different probability distributions for the couplings J_{ij}.

The weak correlations that are present between the couplings in the Hopfield model prevent us from using the same arguments that has led us to (43). In fact, the derivation presented in the chapter XIII of (16) leads to a different result. A similar effect can be observed in the Plefka expansion (52). If the couplings are not simple i.i.d. random variables, the expansion can not be truncated after the second order term. An identification of terms which survive in the limit $N \to \infty$ is necessary (20).

Is there a general way of deriving the correct TAP equations for the different distributions of couplings? The chapters (13) and (18) present different approaches to this problem. The first one is based on identifying new auxiliary variables and couplings between them for which independence is still valid. This leads to TAP like equations which are valid even for a sparse connectivity of couplings. However, the explicit knowledge of the underlying distribution of couplings is required. The second approach motivated by earlier work of (20) develops an adaptive TAP method which does not make explicit assumptions about the distribution. It is however restricted to extensive connectivities.

9 Outlook

We have discussed different types of mean field methods in this chapter. Although we were able to show that in certain limits these approximations become exact, we can not give a general answer to the question how well they will perform on arbitrary real data problems. The situation is perhaps simpler in statistical physics, where there is often more detailed knowledge about the properties of a physical system which helps to motivate a certain approximation scheme. Hence a critical reader may

argue that, especially in cases where MF approaches do not lead to a bound, these approximations are somewhat uncontrolled and can not be trusted. We believe that the situation is less pessimistic. We have seen in this chapter that the MF equations often appear as low order terms in systematic perturbation expansions. Hence, a computation of higher order terms can be useful to check the accuracy of the approximation and may possibly also give error bars on the predictions. We hope that further work in this direction will provide us with approximation methods for complex probabilistic models which are both efficient as well as reliable.

References

[1]S. Amari, S. Ikeda H. Shimokawa, this book.

[2]D. Barber, this book.

[3]H. A. Bethe, Proc. R. Soc. London, Ser A, **151**, 552 (1935).

[4]B. J. Frey and R. Koetter, this book.

[5]Z. Ghahramani and M. J. Beal, this book.

[6]J. J. Hopfield, Proc. Nat. Acad. Sci. USA, **79** 2554 (1982).

[7]K. Humphreys and D. M. Titterington, this book.

[8]Højen-Sørensen, P.A.d.F.R., Winther, O., and Hansen, L. K., Ensemble Learning and Linear Response Theory for ICA, Submitted to NIPS'2000 (2000).

[9]T. Jaakkola, this book.

[10]H. J. Kappen and W. Wiegerinck, this book.

[11]H. J. Kappen and F. B. Rodríguez, Efficient Learning in Boltzmann Machines Using Linear Response Theory, Neural Computation **10**, 1137 (1998).

[12]Y. Kabashima and D. Saad, Belief propagation vs. TAP for decoding corrupted messages, Euro. Phys. Lett. **44**, 668 (1998)

[13]Y. Kabashima and D. Saad, this book.

[14]M. Mézard, The Space of interactions in Neural Networks: Gardner's Computation with the Cavity Method, J. Phys. A (Math. Gen. **22**, 2181 (1989).

[15]M. Mézard and G. Parisi, Mean Field Theory of Randomly Frustrated Systems with Finite Connectivity, Europhys. Lett. **3**, 1067 (1987).

[16]M. Mézard, G. Parisi and M. A. Virasoro, Europhys. Lett. **1**, 77 (1986) and *Spin Glass Theory and Beyond*, Lecture Notes in Physics, 9, World Scientific (1987).

[17]K. Nemoto and H. Takayama, J. Phys. C **18**, L529 (1985).

[18]M. Opper and O. Winther, this book.

[19]G. Parisi, Statistical Field Theory, Addison Wesley, Reading Massachusetts (1988).

[20]G. Parisi and M. Potters, Mean-Field Equations for Spin Models with Orthogonal Interaction Matrices, J. Phys. A (Math. Gen.) **28**, 5267 (1995).

[21]J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann, San Francisco (1988).

[22]F. J. Pineda, C. Resch and I. J. Wang, this book.

[23]Plefka, T., Convergence condition of the TAP equations for the infinite-ranged Ising spin glass model, J. Phys. A **15**, 1971 (1982).

[24]L. K. Saul, T. Jaakkola, M. I. Jordan, Mean Field Theory for Sigmoid Belief Networks, J. Artificial Intelligence Research **4**, 61-76 (1996).

[25]D. Saad, Y. Kabashima and R. Vicente, this book.

[26]D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).

[27]M. Talagrand, Self Averaging and the Space of Interactions in Neural Networks, Random Structures and Algorithms **14**, 199 (1998) and also papers on his webpage `http://www.math.ohio-state.edu/~talagran/`.

[28]T. Tanaka, this book.

[29]D. J. Thouless, P. W. Anderson and R. G. Palmer, Solution of a 'Solvable Model of a Spin Glass', Phil. Mag. **35**, 593 (1977).

[30]Y. Weiss, this book.

[31]K. Y. Wong, S. Li and P. Luo, this book.

[32]J. S. Yedidia, this book.

[33]J. Zinn-Justin, Quantum Field Theory and Critical Phenomena, Clarendon Press, Oxford (1989).