MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

An Idiosyncratic Journey Beyond Mean Field Theory

Jonathan S. Yedidia
MERL
201 Broadway
Cambridge, MA 02139
yedidia@merl.com

TR-2000-27 June 2000

Abstract

I try to clarify the relationships between different ways of deriving or correcting mean field theory, and present "translations" between the language of physicists and that of computer scientists. The connecting thread between the different methods described here is the Gibbs free energy. After introducing the inference problem we are interested in analyzing, I will define the Gibbs free energy, and describe how to derive a mean field approximation to it using a variational approach. I will then explain how one might re-derive and correct the mean field and TAP free energies using high temperature expansions with constrained one-node beliefs. I will explore the relationships between the high-temperature expansion approach, the Bethe approximation, and the belief propagation algorithm, and point out in particular the equivalence of the Bethe approximation and belief propagation. Finally, I will describe Kikuchi approximations to the Gibbs Free energy and advertise new belief propagation algorithms that efficiently compute beliefs equivalent to those obtained from the Kikuchi free energy.

To appear as a chapter in "Advanced Mean Field Methods - Theory and Practice", eds. D. Saad and M. Opper, MIT Press, 2000.

Publication History:–

1. First printing, TR-2000-27, June 2000

1 An Idiosyncratic Journey Beyond Mean Field Theory

Jonathan S. Yedidia

1.1 Introduction

In this chapter I will try to clarify the relationships between different ways of deriving or correcting mean field theory. The December 1999 NIPS workshop on "Advanced Mean Field Methods" succeeded nicely in bringing together physicists and computer scientists, who nowadays often work on precisely the same problems, but come to these problems with different perspectives, methods, names and notations. Some of this chapter is therefore devoted to presenting translations between the language of the physicist and the language of the computer scientist, although I am sure that my original training as a physicist will show through.

I will only cover methods that I have personally used, so this chapter does not attempt to be a thorough survey of its subject. Readers interested in more background on the statistical physics of disordered systems (particularly with regard to the technique of averaging over disorder using the replica method) might also want to consult references (19), (28), and (31), while those interested in the computer science literature on graphical models might consult references (23), (11) and (7).

The connecting thread between the different methods described here is the Gibbs free energy. After introducing the inference problem we are interested in analyzing, I will define the Gibbs free energy, and describe how to derive a mean field approximation to it using a variational approach. I will then explain how one might re-derive and correct the mean field and TAP free energies using high temperature expansions with constrained one-node beliefs. I will explore the relationships between the high-temperature expansion approach, the Bethe approximation, and the belief propagation algorithm, and point out in particular the equivalence of the Bethe approximation and belief propagation. Finally, I will describe Kikuchi approximations to the Gibbs Free energy and advertise new belief propagation algorithms that efficiently compute beliefs equivalent to those obtained from the Kikuchi free energy.

1.2 Inference

We begin by describing the problem we will focus on. In the appealing computer science jargon, this is the problem of "inference." We are given some complicated probabilistic system, which we model by a pair-wise Markov network of N nodes. We label the state of node i by x_i, and write the joint probability distribution function as

$$P(x_1, x_2, ..., x_N) = \frac{1}{Z} \prod_{(ij)} \psi_{ij}(x_i, x_j) \prod_i \psi_i(x_i). \tag{1.1}$$

Here $\psi_{ij}(x_i, x_j)$ is the "compatibility" matrix between connected nodes i and j, $\psi_i(x_i)$ is called the "evidence" for node i, and Z is a normalization constant called the "partition function" by physicists. The notation (ij) means that the sum is over connected nodes.

Such models have many applications, in fields as diverse as computer vision, error-correcting codes, medical diagnosis, and condensed matter physics. It may help your intuition to think of the medical diagnosis application. In such an application, the nodes could represent symptoms and diseases that a patient may have, and the links $\psi_{ij}(x_i, x_j)$ could represent the statistical dependencies between the symptoms and diseases. Note that the links $\psi_{ij}(x_i, x_j)$ would not normally change from one patient to the next. On the other hand, for each patient, we would obtain a different set of evidence $\psi_i(x_i)$, which would correspond to our knowledge of the symptoms for that specific patient. We would like to use the model to infer the probability that the patient has a specific disease–that is, we want to compute a marginal probability like $p_i(x_i)$, which is the probability that the patient has the disease denoted by node i.

I will just give a very rough idea of how such a model might be useful for other applications. In a computer vision application, we might be interested in inferring the shape of an object from the evidence provided by the pixel values of the image. In an error-correcting code, we might be interested in inferring (decoding) the most likely interpretation of a noisy message, where the Markov network itself enforces the error-correcting code. In condensed matter physics, we might want to infer (predict) the response of a magnetic system to the "evidence" of an inhomogeneous magnetic field. For the rest of the chapter, however, I will not make

specific interpretations of the meanings of the nodes, and focus on the mathematics of the problem.

For some networks–small ones or networks that have the topology of a chain or tree–we can compute any desired marginal probabilities exactly, either by explicitly summing over all possible states of the system or by using dynamic programming methods (we will return to the dynamic programming methods, which are also called "belief propagation" algorithms, later in the chapter.) Otherwise, however, we must settle for approximations. If we want to make a distinction between the exact marginal probabilities and approximate ones (something physicists do not usually bother doing explicitly), then we can call the approximation of the exact marginal probability $p_i(x_i)$ the "belief" $b_i(x_i)$, and similarly we call the approximation of the exact two-node marginal probability $p_{ij}(x_i, x_j)$ the belief $b_{ij}(x_i, x_j)$. The mathematical problem we will focus on for the rest of this chapter is as follows: given some arbitrary Markov network defined as in equation (1.1), compute as accurately as possible any desired beliefs.

1.3 Some Models from Statistical Physics

In statistical mechanics, we start with Boltzmann's law for computing joint probability functions:

$$P(x_1, x_2, ..., x_N) = \frac{1}{Z} \exp(-E(x_1, x_2, ..., x_N)/T) \tag{1.2}$$

where E is the energy of the system and T is the temperature. We can re-write equation (1.1) in this way if we define

$$E(x_1, x_2, ..., x_N) = -\sum_{(ij)} J_{ij}(x_i, x_j) - \sum_i h_i(x_i) \tag{1.3}$$

where the "bond strength" function $J_{ij}(x_i, x_j)$ is defined by $\psi_{ij}(x_i, x_j) = \exp(J_{ij}(x_i, x_j)/T)$ and the "magnetic field" $h_i(x_i)$ is defined by $\psi_i(x_i) = \exp(h_i(x_i)/T)$.

Before turning to approximation methods, let us pause to consider some more general and some more specific models. Turning first to more specific models, we can obtain the Ising model by restricting each node i to have two states $s_i = \pm 1$ (for the Ising case, we follow the physics convention and label the states by s_i instead of

x_i), and insisting that the compatibility matrices ψ_{ij} have the form $\psi_{ij} = \begin{pmatrix} \exp(J_{ij}/T) & \exp(-J_{ij}/T) \\ \exp(-J_{ij}/T) & \exp(J_{ij}/T) \end{pmatrix}$ while the evidence vectors have the form $\psi_i = (\exp(h_i/T); \exp(-h_i)/T)$. In that case, we can write the energy as

$$E = -\sum_{(ij)} J_{ij} s_i s_j - \sum_i h_i s_i. \tag{1.4}$$

If we further restrict the J_{ij} to be uniform and positive, we obtain the ferromagnetic Ising model, while if we assume the J_{ij} are chosen from a random distribution, we obtain an Ising spin glass. For these models, the magnetic field h_i is usually, but not always, assumed to be uniform.

We can create more general models by introducing tensors like $\psi_{ijk}(x_i, x_j, x_k)$ in equation (1.1) or equivalently tensors like $J_{ijk}(x_i, x_j, x_k)$ in the energy. One can of course introduce tensors of even higher order. In the extreme limit, one can consider a model where $E(x_1, x_2, ..., x_N) = J_{12...N}(x_1, x_2, ..., x_N)$. If the x_i are binary and the entries of this J tensor are chosen randomly from a Gaussian distribution, we obtain Derrida's Random Energy Model (4).

So far, we have been implicitly assuming that the nodes in the Markov network live on a fixed lattice and that each node can be in a discrete state x_i. In fact, there is nothing to stop us from taking the x_i to be continuous variables, or we can generalize to vectors $\vec{r}_i$, where $\vec{r}_i$ can be interpreted as the position of the ith particle in the system. Looking at it this way, we see that equation (1.3) can be interpreted as an energy function for particles interacting by arbitrary two-body forces in arbitrary one-body potentials.

1.4 The Gibbs Free Energy

Statistical physicists often use the following algorithm when they consider some new model of a physical system:

1. Write down the energy function.
2. Construct an approximate Gibbs free energy.
3. Solve the stationary conditions of the approximate Gibbs free energy.
4. Write paper.

To use this algorithm successfully, one needs to understand what a Gibbs free energy is, and how one might successfully approximate it. We will explore this subject from numerous points of view.

The *exact* Gibbs free energy G_{exact} can be thought of as a mathematical construction designed so that when you minimize it, you will recover Boltzmann's law. G_{exact} is a function of the full joint probability function $P(x_1, x_2, ..., x_N)$ and is defined by

$$G_{exact}(P(x_1, x_2, ..., x_N)) = U - TS \tag{1.5}$$

where U is the average (or "internal") energy:

$$U = \sum_{x_1, x_2, ..., x_N} P(x_1, x_2, ..., x_N) E(x_1, x_2, ..., x_N) \tag{1.6}$$

and S is the entropy:

$$S = -\sum_{x_1, x_2, ..., x_N} P(x_1, x_2, ..., x_N) \ln P(x_1, x_2, ..., x_N). \tag{1.7}$$

If we minimize G_{exact} with respect to $P(x_1, x_2, ..., x_N)$ (one needs to remember to add a Lagrange multiplier to enforce the constraint $\sum_{x_1, x_2, ..., x_N} P(x_1, x_2, ..., x_N) = 1$), we do indeed recover Boltzmann's Law (equation (1.2)) as desired. If we substitute in $P = \exp(-E/T)/Z$ into G_{exact}, we find that at equilibrium (that is, when the joint probability distribution has its correct value), the Gibbs free energy is equal to the Helmholtz free energy defined by $F \equiv -T \ln Z$.

One can understand things this way: the Helmholtz free energy is just a number equal to $U - TS$ at equilibrium, but the Gibbs free energy is a function that gives the value of $U - TS$ when some constraints are applied. In the case of G_{exact}, we constrain the whole joint probablity function $P(x_1, x_2, ..., x_N)$. In other cases that we will look at shortly, we will just constrain some of the marginal probabilities. In general, there can be more than one "Gibbs free energy"–which one you are talking about depends on which additional constraints you want to apply. When we minimize a Gibbs free energy with respect to those probabilities that were constrained, we will obtain self-consistent equations that must be obeyed in equilibrium.

The advantage of working with a Gibbs free energy instead of Boltzmann's Law directly is that it is much easier to come up with ideas for approximations. There are in fact many different approximations

that one could make to a Gibbs free energy, and much of the rest of this chapter is devoted to surveying them.

1.5 Mean Field Theory: The Variational Approach

One very popular way to construct an approximate Gibbs free energy involves a variational argument. The derivation given here will be from a physicist's perspective; for an introduction to variational methods from a different point of view, see (12). Assume that we have some system which can be in, say, K different states. The probability of each state is some number p_α where $\sum_{\alpha=1}^{K} p_\alpha = 1$. Let there be some quantity X_α (like the energy) which depends on which state the system is in, and introduce the notation for the mean value

$$\langle X \rangle \equiv \sum_{\alpha=1}^{K} p_\alpha X_\alpha. \tag{1.8}$$

Then by the convexity of the exponential function, we can prove that

$$\left\langle e^{-X} \right\rangle \geq e^{-\langle X \rangle}. \tag{1.9}$$

Now consider the partition function

$$Z = \sum_\alpha \exp(-E_\alpha/T). \tag{1.10}$$

Let us introduce some arbitrary "trial" energy function E^0_α. We can manipulate Z into the form

$$Z = \frac{\sum_\alpha \exp(-(E_\alpha - E^0_\alpha)/T) \exp(-E^0_\alpha/T)}{\sum_\alpha \exp(-E^0_\alpha/T)} \sum_\alpha \exp(-E^0_\alpha/T) \tag{1.11}$$

or

$$Z = \left\langle e^{-(E-E^0)/T} \right\rangle_0 \sum_\alpha \exp(-E^0_\alpha/T) \tag{1.12}$$

where the notation $\langle X \rangle_0$ means the average of X_α using a trial probability distribution

$$p^0_\alpha = \frac{\exp(-E^0_\alpha/T)}{\sum_\alpha \exp(-E^0_\alpha/T)}. \tag{1.13}$$

We can now use the inequality (1.9) to assert that

$$Z \geq e^{-\left\langle (E-E^0)/T \right\rangle_0} \sum_\alpha \exp(-E^0_\alpha/T) \tag{1.14}$$

for *any* function E^0_α. In terms of the Helmholtz free energy $F \equiv -T \ln Z$, we can equivalently assert that

$$F \leq -T \ln \sum_\alpha \exp(-E^0_\alpha/T) + \left\langle E - E^0 \right\rangle_0 \equiv F_{var} \tag{1.15}$$

where we define the quantity on the right-hand side of the inequality as the variational mean field free energy F_{var} corresponding to the trial probability function p^0_α. A little more manipulation gives us

$$F_{var} = \langle E \rangle_0 - T S_0 \geq F \tag{1.16}$$

where S_0 is the trial entropy defined by $S_0 = -\sum_\alpha p^0_\alpha \ln p^0_\alpha$. This inequality gives us a useful variational argument: we will look for the trial probability function p^0_α which gives us the lowest variational free energy.

To be able to use the variational principle in practice, we must restrict ourselves to a class of probabilities for which we can actually analytically compute F_{var}. The quality of the variational approximation will depend on how well the trial probability function can represent the true one. For continuous x_i or $\vec{r}_i$, one can use Gaussians as very good, yet tractable variational functions (28; 2; 3). Richard Feynman was one of the first physicists to use this kind of variational argument (with Gaussian trial probability functions) in his treatment of the polaron problem (5).

The variational probability functions that are tractable for discrete x_i are not nearly as good. When people talk about "mean field theory," they are usually referring to using a trial probability function of the factorized form

$$p^0(x_1, x_2, ..., x_N) = \prod_i b_i(x_i). \tag{1.17}$$

and computing F_{var} for some energy function of a form like equation (1.3). The "mean field" Gibbs free energy that results is

$$G_{MF} = -\sum_{(ij)} \sum_{x_i, x_j} J_{ij}(x_i, x_j) b_i(x_i) b_j(x_j) - \sum_i \sum_{x_i} h_i(x_i) b_i(x_i)$$

$$+T \sum_i \sum_{x_i} b_i(x_i) \ln b_i(x_i) \tag{1.18}$$

To obtain the beliefs in equilibrium according to this approximation, one minimizes G_{MF} with respect to the beliefs $b_i(x_i)$. Let us see how this works for the Ising model with no external field. In that case, it makes sense to define the local magnetization

$$m_i \equiv \langle s_i \rangle = b_i(s_i = 1) - b_i(s_i = -1) \tag{1.19}$$

which is a scalar that can take on values from -1 to 1. In terms of the magnetization, we have

$$\begin{aligned} G_{MF} &= -\sum_{(ij)} J_{ij} m_i m_j \\ &\quad +T \sum_i \left[\frac{1+m_i}{2} \ln \left(\frac{1+m_i}{2} \right) + \frac{1-m_i}{2} \ln \left(\frac{1-m_i}{2} \right) \right] \end{aligned} \tag{1.20}$$

and the mean field stationary conditions are

$$m_i = \tanh \left(\frac{\sum_j J_{ij} m_j}{T} \right). \tag{1.21}$$

If we further specialize to the case of a ferromagnet on a d-dimensional hyper-cubic lattice, set all the $J_{ij} = \frac{1}{2d}$, and assume that all m_i are equal to the same magnetization m, we can analytically analyze the solutions of this equation. We find that above $T_c = 1$, the only solution is $m = 0$, while below T_c, we have two other solutions with positive or negative magnetization. This is a classic example of a phase transition that breaks the underlying symmetry in a model. The mean field prediction of a phase transition is qualitatively correct for dimension $d \geq 2$. Other bulk thermodynamic quantities like the susceptibility $\chi \equiv \partial m / \partial h$ and the specific heat $C \equiv \partial U / \partial T$ are also easy to compute once we have the stationary conditions.

How good an approximation does mean field theory give? It depends a lot on the model. For the Ising ferromagnet, mean field theory becomes exact for a hyper-cubic lattice in the limit of infinite dimensions, or for an "infinite-ranged" lattice where every node is connected to every other node. On the other hand, for lower dimensional ferromagnets, or spin glasses in any dimension, mean field theory can give quite poor results. In general, mean field theory does badly when the nodes in a network

fluctuate a lot around their mean values, because it incorrectly insists that all two-node beliefs $b_{ij}(x_i, x_j)$ are simply given by $b_{ij}(x_i, x_j) = b_i(x_i)b_j(x_j)$. In practice, one sees many papers where questionable mean field approximations are used when it would not have been too difficult to obtain better results using one of the techniques that I describe in the rest of the chapter.

1.6 Correcting Mean Field Theory

Mean field theory is exact for the infinite-ranged ferromagnet, so when physicists started contemplating spin glasses in the 1970's, they quickly turned to the simplest corresponding model: the infinite-ranged Sherrington-Kirpatrick (SK) Ising spin glass model with zero field and J_{ij}'s chosen from a zero-mean Gaussian distribution (25). Thouless, Anderson and Palmer (TAP) presented "as a *fait accompli*" (26) a Gibbs free energy that they claimed should be exact for this model:

$$\begin{aligned} -\beta G_{TAP} &= -\sum_i \left[\frac{1+m_i}{2} \ln\left(\frac{1+m_i}{2}\right) + \frac{1-m_i}{2} \ln\left(\frac{1-m_i}{2}\right) \right] \\ &\quad +\beta \sum_{(ij)} J_{ij} m_i m_j + \frac{\beta^2}{2} \sum_{(ij)} J_{ij}^2 (1-m_i^2)(1-m_j^2) \end{aligned} \tag{1.22}$$

where $\beta \equiv 1/T$ is the inverse temperature. The only difference between the TAP and ordinary mean field free energy is the last term, which is sometimes called the "Onsager reaction" term.

I have written the TAP free energy in a suggestive form: it appears to be a Taylor expansion in powers of β. Plefka showed that one could in fact derive G_{TAP} from such a Taylor expansion (24). Antoine Georges and I later (10) showed how to continue the Taylor expansion to terms beyond $O(\beta^2)$, and exploited this kind of expansion for a variety of statistical mechanical (8; 30) and quantum mechanical (9) models. Of course, the higher-order terms are important for any model that is not infinite-ranged. Because this technique is little-known, but quite generally applicable, I will review it here using the Ising spin glass energy function.

The variational approximation gives a rigorous upper bound on the Helmholtz free energy, but there is no reason to believe that it is the best approximation one can make for the magnetization-dependent Gibbs free

energy. We can construct such a Gibbs free energy by adding a set of external auxilary fields (Lagrange multipliers) that are used to insure that all the magnetizations are constrained to their desired values. Note that the auxiliary fields are temperature-dependent. Of course, when the magnetizations are at their *equilibrium* values, no auxiliary fields will be necessary. We write

$$-\beta G = \ln \sum_{s_1,..,s_N} \exp\left(\beta \sum_{(ij)} J_{ij} s_i s_j + \sum_i \lambda_i(\beta)(s_i - m_i)\right) \tag{1.23}$$

where the $\lambda(\beta)$ are our auxiliary fields.

We can use this exact formula to expand $-\beta G(\beta, m_i)$ around $\beta = 0$:

$$-\beta G = -(\beta G)_{\beta=0} - \left(\frac{\partial(\beta G)}{\partial \beta}\right)_{\beta=0} \beta - \left(\frac{\partial^2(\beta G)}{\partial \beta^2}\right)_{\beta=0} \frac{\beta^2}{2} - ... \tag{1.24}$$

At $\beta = 0$, the spins are entirely controlled by their auxiliary fields, and so we have reduced our problem to one of independent spins. Since m_i is fixed equal to $\langle s_i \rangle$ for any inverse temperature β, it is in particular equal to $\langle s_i \rangle$ when $\beta = 0$, which gives us the relation

$$m_i = \langle s_i \rangle_{\beta=0} = \frac{\sum_{s_i=\pm 1} s_i \exp(\lambda_i(0) s_i)}{\sum_{s_i=\pm 1} \exp(\lambda_i(0) s_i)} = \tanh(\lambda_i(0)) \tag{1.25}$$

From the definition of $-\beta G(\beta, m_i)$ given in equation (1.23), we find that

$$-(\beta G)_{\beta=0} = \sum_i \ln\left[\cosh(\lambda_i(0))\right] - \lambda_i(0) m_i. \tag{1.26}$$

Eliminating the $\lambda_i(0)$, we obtain

$$-(\beta G)_{\beta=0} = -\sum_i \left[\frac{1+m_i}{2} \ln\left(\frac{1+m_i}{2}\right) + \frac{1-m_i}{2} \ln\left(\frac{1-m_i}{2}\right)\right] \tag{1.27}$$

which is just the mean field entropy. Considering next the first derivative, we find that

$$-\beta\left(\frac{\partial(\beta G)}{\partial \beta}\right)_{\beta=0} = \beta \left\langle \sum_{(ij)} J_{ij} s_i s_j \right\rangle_{\beta=0} + \beta \langle s_i - m_i \rangle_{\beta=0} \left.\frac{\partial \lambda_i}{\partial \beta}\right|_{\beta=0}. \tag{1.28}$$

At $\beta = 0$, the two-node correlation functions factorize so we find that

$$-\beta\left(\frac{\partial(\beta G)}{\partial\beta}\right)_{\beta=0} = \beta\sum_{(ij)} J_{ij}m_i m_j \tag{1.29}$$

which is, of course, the same as the variational internal energy term.

Naturally, we can continue this expansion to arbitrarily high order if we work hard enough. Unfortunately, neither Georges and I, nor Parisi and Potters who later examined this expansion (22), were able to derive the Feynman rules for a fully diagrammatic expansion, but there are some tricks that make the computation easier (10). To order β^4, we find that

$$\begin{aligned}
-\beta G \;=\; & -\sum_i \left[\frac{1+m_i}{2}\ln\left(\frac{1+m_i}{2}\right) + \frac{1-m_i}{2}\ln\left(\frac{1-m_i}{2}\right)\right] \\
& +\beta\sum_{(ij)} J_{ij}m_i m_j \\
& +\frac{\beta^2}{2}\sum_{(ij)} J_{ij}^2(1-m_i^2)(1-m_j^2) \\
& +\frac{2\beta^3}{3}\sum_{(ij)} J_{ij}^3 m_i(1-m_i^2)m_j(1-m_j^2) \\
& +\beta^3\sum_{(ijk)} J_{ij}J_{jk}J_{ki}(1-m_i^2)(1-m_j^2)(1-m_k^2) \\
& -\frac{\beta^4}{12}\sum_{(ij)} J_{ij}^4(1-m_i^2)(1-m_j^2)(1+3m_i^2+3m_j^2-15m_i^2m_j^2) \\
& +2\beta^4\sum_{(ijk)} J_{ij}^2J_{jk}J_{ki}m_i(1-m_i^2)m_j(1-m_j^2)(1-m_k^2) \\
& +\beta^4\sum_{(ijkl)} J_{ij}J_{jk}J_{kl}J_{li}(1-m_i^2)(1-m_j^2)(1-m_k^2)(1-m_l^2) \\
& +\ldots
\end{aligned} \tag{1.30}$$

where the notation (ij), (ijk), or $(ijkl)$ means that one should sum over all distinct pairs, triplets, or quadruplets of spins.

For the ferromagnet on a d-dimensional hypercubic lattice, all these terms can be reorganized according to their contribution in powers of $1/d$. It is easy to show that only the mean field terms contribute in

the limit $d \to \infty$ and to generate $1/d$ expansions for all the bulk thermodynamic quantities, including the magnetization (10).

A few points should be made about the Taylor expansion of equation (1.30). First, as with any Taylor expansion, there is a danger that the radius of convergence of the expansion will be too small to obtain results for the value of β you are interested in. It is hard to say anything about this issue in general. For ferromagnets, there does not seem to be any problem at low or high temperatures, but for the SK model, the issue is non-trivial and was analyzed by Plefka (24).

Secondly, since the expansion was presented as one that starts at $\beta = 0$, it is initially surprising that it can work at low temperatures. The explanation, at least for the ferromagnetic case, is that the higher-order terms become exponentially small in the limit $T \to 0$. Thus, the expansion works very well for $T \to 0$ or $T \to \infty$ and is worst near T_c.

Finally, the TAP free energy is sometimes justified as a "Bethe approximation," that is, as an approximation that would become exact on a tree-like lattice (1). In fact, the general convention in the statistical physics community is to refer to the technique of using a Bethe approximation on a inhomogeneous model as the "TAP approach." In general, to obtain the proper Bethe approximation from the expansion (1.30) for models on a tree-like lattice, we need to sum over all the higher-order terms that do not include loops of nodes. The TAP free energy for the SK model only simplifies because for that model all terms of order β^3 or higher are believed to vanish anyways in the limit $N \to \infty$ (which is the "thermodynamic limit" physicists are interested in). In the next section, we will describe a much simpler way to arrive at the important Bethe approximation.

1.7 The Bethe Approximation

The remaining sections of this chapter will discuss the Bethe and Kikuchi approximations and belief propagation algorithms. My understanding of these subjects was formed by a collaboration with Bill Freeman at MERL and Yair Weiss at Berkeley. These sections can be considered an introduction to the work that we did together (29).

So far we have discussed Gibbs free energies with just one-node beliefs $b_i(x_i)$ constrained. The next obvious step to take is to constrain

the two-node beliefs $b_{ij}(x_i, x_j)$ as well. For Markov networks that have a tree-like topology, taking this step is sufficient to obtain the exact Gibbs free energy. The reason is that for these models, the exact joint probability distribution itself can be factorized into a form that only depends on one-node and two-node marginal probabilities:

$$p(x_1, x_2, ..., x_N) = \prod_{(ij)} p_{ij}(x_i, x_j) \prod_i [p_i(x_i)]^{1-q_i} \tag{1.31}$$

where q_i is the number of nodes that are connected to node i.

Recall that the exact Gibbs free energy is $G = U - TS$, where the internal energy $U = \sum_\alpha p_\alpha E_\alpha$, the entropy $S = \sum_\alpha p_\alpha \ln p_\alpha$, and α is an index over every possible state. Using equation (1.31), we find that the exact entropy for models with tree-like topology is

$$\begin{aligned} S &= -\sum_{(ij)} \sum_{x_i, x_j} p_{ij}(x_i, x_j) \ln p_{ij}(x_i, x_j) \\ &\quad - \sum_i (1 - q_i) \sum_{x_i} p_i(x_i) \ln p_i(x_i). \end{aligned} \tag{1.32}$$

The average energy can be expressed exactly in terms of one-node and two-node marginal probabilities for pair-wise Markov networks of *any* topology:

$$\begin{aligned} U &= -\sum_{(ij)} \sum_{x_i, x_j} p_{ij}(x_i, x_j)(J_{ij}(x_i, x, j) + h_i(x_i) + h_j(x_j)) \\ &\quad - \sum_i (1 - q_i) \sum_{x_i} p_i(x_i) h_i(x_i). \end{aligned} \tag{1.33}$$

The first term is just the average energy of each link, and the second term is a correction for the fact that the evidence at each node is counted $q_i - 1$ times too many.

The Bethe approximation to the Gibbs free energy amounts to using these expressions (with beliefs substituting for exact marginal probabilities) for any pair-wise Markov network:

$$\begin{aligned} G_{Bethe}(b_i, b_{ij}) &= \sum_{(ij)} \sum_{x_i, x_j} b_{ij}(x_i, x_j)(T \ln b_{ij}(x_i, x_j) + E_{ij}(x_i, x_j)) \\ &\quad + \sum_i (1 - q_i) \sum_{x_i} b_i(x_i)(T \ln b_i(x_i) + E_i(x_i)) \end{aligned} \tag{1.34}$$

where we have introduced the local energies $E_i(x_i) \equiv -h_i(x_i)$ and

$E_{ij}(x_i, x_j) \equiv -J_{ij}(x_i, x_j) - h_i(x_i) - h_j(x_j)$. Of course, the beliefs $b_{ij}(x_i, x_j)$ and $b_i(x_i)$ must obey the standard normalization conditions $\sum_{x_i} b_i(x_i) = 1$ and $\sum_{ij} b_{ij}(x_i, x_j) = 1$ and marginalization conditions $b_i(x_i) = \sum_{x_j} b_{ij}(x_i, x_j)$.

There is more than one way to obtain the stationarity conditions for the Bethe free energy. For inhomogeneous models, the most straightforward approach is to form a Lagrangian L by adding Lagrange multipliers which enforce the normalization and marginalization conditions and to differentiate the Lagrangian with respect to the beliefs and those Lagrange multipliers. We have

$$\begin{aligned} L &= G_{Bethe} + \sum_{(ij)} \sum_{x_j} \lambda_{ij}(x_j) \left(b_j(x_j) - \sum_{x_i} b_{ij}(x_i, x_j) \right) \\ &+ \sum_{(ij)} \sum_{x_i} \lambda_{ji}(x_i) \left(b_i(x_i) - \sum_{x_j} b_{ij}(x_i, x_j) \right) \\ &+ \sum_i \gamma_i \left(1 - \sum_{x_i} b_i(x_i) \right) + \sum_{(ij)} \gamma_{ij} \left(1 - \sum_{x_i, x_j} b_{ij}(x_i, x_j) \right) \end{aligned} \tag{1.35}$$

Of course, the derivatives with respect to the Lagrange multipliers give back the desired constraints, while the derivatives with respect to the beliefs give back equations for beliefs in terms of Lagrange multipliers:

$$b_i(x_i) = \frac{1}{Z_i} \exp \left[-\frac{E_i(x_i)}{T} + \frac{\sum_j \lambda_{ji}(x_i)}{T(q_i - 1)} \right] \tag{1.36}$$

and

$$b_{ij}(x_i, x_j) = \frac{1}{Z_{ij}} \exp \left[-\frac{E_{ij}(x_i, x_j)}{T} + \frac{\lambda_{ji}(x_i)}{T} + \frac{\lambda_{ij}(x_j)}{T} \right] \tag{1.37}$$

where Z_i and Z_{ij} are constants which enforce the normalization conditions. Finally one can use the marginalization conditions to obtain self-consistent equations for the Lagrange multipliers.

The Bethe approximation is a significantly better approximation to the Gibbs free energy than the mean field approximation. The only real difficulty is a practical one: how do we minimize the Bethe free energy efficiently? As we shall see, it turns out that the belief propagation algorithm, which was developed by Pearl following an entirely different

path, provides a possible answer.

1.8 Belief Propagation

Belief propagation algorithms can probably best be understood by imagining that each node in a Markov network represents a person, who communicates by "messages" with those people on connected nodes about what their beliefs should be. Let us see what the properties of these messages should be if we want to get reasonable equations for the beliefs $b_i(x_i)$. We will denote the message from node j to node i by $M_{ji}(x_i)$. Note that the message has the same dimensionality as node i–the person at j is telling the one at i something like "you should believe in your state 1 twice as strongly as your state 2, and your state number 3 should be impossible." That message would be the vector $(2, 1, 0)$. Now imagine that the person at node i is looking at all the messages that he is getting, plus the independent evidence that he alone is receiving denoted by $\psi_i(x_i)$. Assume that each message is arriving independently and is reliably informing the person at node i about something he has no other way of finding out. Given equally reliable messages and evidence, what should his beliefs be? A reasonable guess would be

$$b_i(x_i) = \alpha\psi_i(x_i) \prod_{j \in N(i)} M_{ji}(x_i) \tag{1.38}$$

where α is a normalization constant, and $N(i)$ denotes all the nodes neighboring i. Thus a person following this rule who got messages $(2, 1, 0)$ and $(1, 1, 1)$ and had personal evidence $(1, 2, 1)$ would have a belief $(.5, .5, 0)$. His thought process would work like this: "The first message is telling me that state 3 is impossible, the second message can be ignored because it is telling me it does not care, while my personal evidence is telling me to believe in state 2 twice as strongly as state 1, which is the opposite of what the first message tells me, so I will just believe in state 1 and state 2 equally strongly."

Now consider the joint beliefs of a pair of neighboring nodes i and j. Clearly they must depend on the compatibility matrix $\psi_{ij}(x_i, x_j)$, the evidence at each node $\psi_i(x_i)$ and $\psi_j(x_j)$, and all the messages coming

into nodes i and j. The obvious guess would be the rule

$$b_{ij}(x_i, x_j) = \alpha \psi_{ij}(x_i, x_j) \psi_i(x_i) \psi_j(x_j) \prod_{k \in N(i)} M_{ki}(x_i) \prod_{l \in N(j)} M_{lj}(x_j) \quad (1.39)$$

If we combine these rules for the one-node and two-node beliefs with the marginalization condition

$$b_i(x_i) = \sum_{x_j} b_{ij}(x_i, x_j) \quad (1.40)$$

we obtain the self-consistent equations for the messages

$$M_{ij}(x_j) = \alpha \sum_{x_i} \psi_{ij}(x_i, x_j) \psi_i(x_i) \prod_{k \in N(i) \backslash j} M_{ki}(x_i) \quad (1.41)$$

where $N(i) \backslash j$ means all nodes neighboring i except for j. The belief propagation algorithm amounts to solving these message equations iteratively, and using the solution for the messages in the belief equations.

So far I have probably just convinced you that the belief propagation algorithm is vaguely plausible. Pearl did more than that of course–he showed directly that all the belief propagation equations written above are exact for Markov networks that have a tree-like topology (23). One might note that this fact was already partially known in the physics literature–as long ago as 1979, T. Morita wrote down the correct belief propagation equations for the case of an Ising spin glass in a random field (20). Of course, the suitability of these equations as an algorithm was not appreciated. Recently, Y. Kabashima and D. Saad (13; 14) have shown that for a number of other specific disordered models, the TAP approach and belief propagation give rise to identical equations, and speculated that this might be true in general.

Freeman, Weiss and I have shown that this identity does in fact hold in general (29). To prove it for general Markov networks, you simply need to identify the following relationship between the Lagrange multipliers $\lambda_{ij}(x_j)$ that we introduced in the last section and the messages $M_{ij}(x_j)$:

$$\lambda_{ij}(x_j) = T \ln \prod_{k \in N(j) \backslash i} M_{kj}(x_j) \quad (1.42)$$

Using this relation, one can easily show that equations (1.36) and (1.37) derived for the Bethe approximation in the last section are equivalent to the belief propagation equations (1.38) and (1.39).

1.9 Kikuchi Approximations and Generalized Belief Propagation

Pearl pointed out that belief propagation was not exact for networks with loops, but that has not stopped a number of researchers from using it on such networks, often very successfully. One particularly dramatic case is near Shannon-limit performance of "Turbo codes" and low density parity check codes, whose decoding algorithm is equivalent to belief propagation on a network with loops (18; 17). For some problems in computer vision involving networks with loops, belief propagation has worked well and converged very quickly (7; 6; 21). On the other hand, for other networks with loops, belief propagation gives poor results or fails to converge (21; 29).

What has been generally missing has been an idea for how one might systematically correct belief propagation in a way that preserves its main advantage–the rapidity with which it normally converges (27). The idea which turned out to be successful was to work out approximations to the Gibbs free energy that are even more accurate than the Bethe approximation, and find corresponding "generalized" belief propagation algorithms.

Once one has the idea of improving the approximation for the Gibbs free energy by constraining two-node beliefs like $b_{ij}(x_i, x_j)$, it is natural to go further and constrain higher-order beliefs as well. The "cluster variation method," which was invented by Kikuchi (15; 16), is a way of obtaining increasingly accurate approximations in precisely this way. The idea is to group the nodes of the Markov network into basic (possibly overlapping) clusters, and then to compute an approximation to the Gibbs free energy by summing the free energies of the basic clusters, minus the free energy of over-counted intersections of clusters, minus the free energy of over-counted intersections of intersections, and so on. The Bethe approximation is the simplest example of one of these more complicated Kikuchi free energies: for that case, the basic clusters are all the connected pairs of nodes. Every Kikuchi free energy will handle the average energy exactly, and the entropy will become increaingly accurate as the size of the basic clusters increases.

Rather than repeat analysis that you can find elsewhere, I will just advertise the results of our work (29). One can indeed derive new belief propagation algorithms based on Kikuchi free energies. They converge to beliefs that are provably equivalent to the beliefs that are obtained from

the Kikuchi stationary conditions. The new messages that need to be introduced involve groups of nodes telling other groups of nodes what their joint beliefs should be. These new belief propagation algorithms have the attractive feature of being user-adjustable: by paying some additional computational cost, you can buy additional accuracy. In practice, the additional cost is not great: we found that we were able to obtain dramatic improvements in accuracy at negligible cost for some models where ordinary belief propation performs poorly.

Acknowledgements

It is a pleasure to thank my collaborators Jean-Philippe Bouchaud, Bill Freeman, Antoine Georges, Marc Mézard, and Yair Weiss with whom I have enjoyed exploring the issues described in this chapter.

References

[1]Bethe, H.A. 1935, Proc. Roy. Soc. London A 150, 552

[2]Bouchaud, J.P., Mézard, M., Parisi, G, and Yedidia, J.S. 1991, J. Phys. A. 24, L1025

[3]Bouchaud, J.P., Mézard, M., and Yedidia, J.S. 1992, Phys. Rev. B 46, 14686

[4]Derrida, B. 1981, Phys. Rev. B. 24, 2613

[5]Feynman, R.P. 1955, Phys. Rev. 97, 660

[6]Freeman, W.T., and Pasztor, E. 1999, 7th Intl. Conf. Computer Vision, 1182

[7]Frey, B.J 1998, Graphical Models for Machine Learning and Digital Communication, Cambridge: MIT Press

[8]Georges, A., Mézard M., and Yedidia, J.S. 1990, Phys. Rev. Lett. 64, 2937

[9]Georges, A. and Yedidia, J.S. 1991, Phys. Rev B 43, 3475

[10]Georges, A. and Yedidia, J.S. 1991, J. Phys. A 24, 2173

[11]Jordan, M.I., ed. 1998, Learning in Graphical Models, Cambridge: MIT Press

[12]Jordan, M.I., Ghahramani, Z., Jaakola, T., and Saul, L.K. 1998, Learning in Graphical Models, M.I. Jordan ed., Cambridge: MIT Press

[13]Kabashima, Y. and Saad, D. 1998, Euro. Phys. Lett. 44, 668

[14]Kabashima, Y. and Saad, D. 2000, Contribution to this volume

[15]Kikuchi, R. 1951, Phys. Rev. 81, 988

[16]Kikuchi, R. 1994, Special issue in honor of R. Kikuchi, Prog. Theor. Phys. Suppl., 115

[17]MacKay, D.J.C. 1999, IEEE Trans. on Inf. Theory

[18]McEliece, R. MacKay, D.J.C., and Cheng, J. 1998, IEEE J. on Sel Areas in Comm. 16(2), 140

[19]Mézard, M., Parisi, G., and Virasoro, M.A. 1987, Spin Glass Theory and Beyond, Singapore: World Scientific

[20]Morita, T. 1979, Physica 98A, 566

[21]Murphy, K., Weiss, Y., and Jordan, M. 1999, in Proc. Uncertainty in AI.

[22]Parisi, G. and Potters, M. 1995, J. Phys. A 28, 5267

[23]Pearl, J. 1988, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, San Francisco: Morgan Kaufman

[24]Plefka, T. 1982, J. Phys. A 15, 1971

[25]Sherrington, D. and Kirpatrick, S. 1975, Phys. Rev. Lett. 35, 1792

[26]Thouless, D.J., Anderson, P.W., and Palmer, R.G. 1977, Phil. Mag. 35, 593

[27]Weiss, Y. 1999, Bayesian Belief Propagation for Image Understanding, available at Yair Weiss's homepage

[28]Yedidia, J.S. 1993, 1992 Lectures in Complex Systems, L. Nadel and D. Stein, eds., Addison-Wesley, 299

[29]Yedidia, J.S, Freeman, W.T., and Weiss, Y. 2000, MERL TR2000-26 available at http://www.merl.com/reports/TR2000-26/

[30]Yedidia, J.S. and Georges, A. 1990, J. Phys. A 23, 2165

[31]Young, A.P., ed. 1998, Spin Glasses and Random Fields, World Scientific