1 Adaptive TAP Equations

Manfred Opper and Ole Winther

We develop a TAP mean field approach to models with quadratic interactions which does not assume a specific randomness of the couplings but rather adapts to the concrete data. The method is based on an extra set of mean field equations for the Onsager correction term to the naive mean field result. We present applications for the Hopfield model and for a Bayesian classifier.

1.1 Introduction

Mean field (MF) methods provide efficient approximations which are able to cope with the increasing complexity of modern probabilistic data models. They replace the intractable task of computing high dimensional sums and integrals by the tractable problem of solving a system of nonlinear equations.

The TAP (21) MF approach represents a principled way for correcting the deficiencies of simple MF methods which are based on the crude approximation of replacing the intractable distribution by a factorized one, thereby neglecting important correlations between variables. In contrast, the TAP method takes into account nontrivial dependencies by estimating the reaction of all other random variables when a single variable is deleted from the system (8).

The method has its origin in the statistical physics of amorphous systems, where it was developed by Thouless, Anderson and Palmer (TAP) to treat the Sherrington-Kirkpatrick (SK) model of disordered magnetic materials (19). Under the assumption that the couplings (or interactions) between random variables are themselves drawn at random from certain classes of distributions, the TAP equations provide an *exact* result in the 'thermodynamic limit' of infinitely many variables.

The 'Onsager correction' to the simple or 'naive' MF theory will explicitly depend on the distribution of these couplings. Two models with the same connectivities but different distributions for the couplings, like e.g. the SK model and the Hopfield model (5) have different expressions for the TAP corrections (see e.g. (8), chapter XIII).

In order to use the TAP method as a good approximation for practical applications to real data, the lack of knowledge of the underlying distribution of the couplings should be compensated by an algorithm which *adapts* the Onsager correction to the *concrete* set of couplings. Simply taking the correction from a theory that *assumes* a specific distribution may lead to suboptimal performance.

It is the goal of this chapter to introduce such an adaptive TAP scheme which has been motivated by work of (15) who derived TAP equations for models with non-iid distributions of couplings. Our method generalizes our previous papers (12; 13), which were devoted to specific Gaussian process applications, to general models with quadratic interactions. When applied to the 'thermodynamic' limit of fully connected models with specified distributions, our method reproduces the known exact results. It differs however from TAP approaches which are based on second order expansions (17) of the Gibbs free energy with respect to small couplings. Our approach usually contains contributions from all orders in the perturbation theory (15).

C++ software that implements the TAP, 'naive' mean field and linear response algorithms for a number of different models with quadratic interactions is available at `http://www.thep.lu.se/tf2/staff/winther`.

The rest of this chapter is organized as follows. Section 1.2 defines the models with quadratic interactions. In Section 1.3, we present the basic derivation of our adaptive TAP equations. In section 1.4, we give a recipe for solving the mean field equations. Sections 1.5 and 1.6 present two examples of the approach. An outlook is given in 1.7.

1.2 Models with quadratic interactions

In this chapter we study models defined by distributions of the type

$$P(\mathbf{S}) = \frac{1}{Z} \prod_j \rho_j(S_j) \ \exp\left[\sum_{i<j} S_i J_{ij} S_j + \sum_i S_i \theta_i\right] \ . \tag{1.1}$$

$\mathbf{S} = (S_1, \ldots, S_N)$ is a vector of random variables and $\rho_j(S)$ denotes a single variable distribution in which we also encode all the constraints to be satisfied by S_j (discreteness, bounds on the range etc.). θ_i is an external field. The partition function is given by

$$Z = \int \prod_j (dS_j \rho(S_j)) \, \exp\left[\sum_{i<j} S_i J_{ij} S_j + \sum_i S_i \theta_i\right] \, . \tag{1.2}$$

This type of distribution can be used to describe a variety of important probabilistic models. We will only mention a few.

Setting $\rho(S_i) = \frac{1}{2}\delta(S_i - 1) + \frac{1}{2}\delta(S_i + 1)$, the S_i become Ising spin variables and the model is of the Boltzmann machine type (1; 16; 6; 20).

For the second example, we assume that independent stochastic observations y are modelled by a likelihood $P(y|S)$ where S is a latent variable defined on a space of input variables $\mathbf{x}$. We further assume that the prior distribution of a vector $\mathbf{S} = (S(\mathbf{x}_1), \ldots, S(\mathbf{x}_N))$ of latent variables for N spatial inputs is a joint Gaussian with zero means and covariance $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, where K is a positive definite kernel. In other words, $\mathbf{S}$ is a sample of N points from a Gaussian process. If we set $J_{ij} = -(K^{-1})_{ij}$ for $j \neq i$ and $\rho_i(S_i) \propto e^{-\frac{1}{2}S_i^2 (K^{-1})_{ii}} P(y_i|S_i)$, then (1.1) is just the posterior distribution for $\mathbf{S}$.

(1.1) can also be used for the *Independent Component Analysis* (7; 4). The classes of models to which (1.1) applies can be expanded further by freely allowing that certain variables are used outside of their 'natural' range. E.g., when the S_i are extended to n dimensional vectors, a suitable continuation of $n \to 0$ applies to the *Matching Problem*, which is a combinatorial optimization problem studied in (9). A further example is given later in this chapter where we allow the variables $\mathbf{S}$ to be extended to the complex plane.

The aims of a mean field theory for the model (1.1) are to compute approximate values for the moments of the variables S_i and and/or for the free energy $-\ln Z$. In the following we will restrict ourselves to mean field equations for the first and second moments of $\mathbf{S}$ i.e. for $\langle S_i \rangle$ and $\chi_{ij} \equiv \langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle$, where we denote an average with respect to $P(\mathbf{S})$ by $\langle \cdot \rangle$. For a discussion of approximations to $-\ln Z$ within different mean field approaches, see e.g. Refs. (8; 2).

1.3 Deriving the TAP equations

We derive TAP equations using the cavity method (8; 11; 13). The starting point is the following exact equation for the marginal distribution of the variable S_i

$$P_i(S_i) = \int \prod_{j,j\neq i} dS_j \; P(\mathbf{S}) = \frac{\int \prod_{j,j\neq i} dS_j \; \rho(S_i) \; e^{S_i(\sum_j J_{ij} S_j + \theta_i)} \; P(\mathbf{S}\backslash S_i)}{\int \prod_j dS_j \; \rho(S_i) \; e^{S_i(\sum_j J_{ij} S_j + \theta_i)} \; P(\mathbf{S}\backslash S_i)} \;. \tag{1.3}$$

$P(\mathbf{S}\backslash S_i)$ is the distribution of all variables excluding S_i in a system where the ith variable is absent. [1] We see from eq. (1.3) that S_i interacts with the remaining variables only through the field $h_i = \sum_j J_{ij} S_j$. Hence, we introduce its 'cavity' distribution, i.e. the distribution of the field at the 'position' of the 'empty' site i by

$$P(h_i\backslash S_i) = \int \prod_{j\neq i} dS_j \; \delta\left(h_i - \sum_j J_{ij} S_j\right) P(\mathbf{S}\backslash S_i) \tag{1.4}$$

and rewrite (1.3) as

$$P_i(S_i) = \frac{\rho(S_i) \; \left\langle e^{S_i(h_i+\theta_i)}\right\rangle_{\backslash i}}{\left\langle \int dS_i \; \rho(S_i) \; e^{S_i(h_i+\theta_i)}\right\rangle_{\backslash i}} \;, \tag{1.5}$$

where $\langle\cdot\rangle_{\backslash i}$ denotes the average with respect to the cavity distribution. In general, an exact computation of the cavity distribution is formidable, so we have to resort to approximations.

The first approximation is based on the assumption of weak dependencies between the random variables S_j. Expecting that in such a case an appropriate central limit theorem justifies the approximation of (1.4) by a Gaussian distribution, we set

$$P(h_i\backslash S_i) \approx \frac{1}{\sqrt{2\pi V_i}} \exp\left(-\frac{(h_i - \langle h_i\rangle_{\backslash i})^2}{2V_i}\right) \;, \tag{1.6}$$

where $V_i \equiv \langle h_i^2\rangle_{\backslash i} - \langle h_i\rangle_{\backslash i}^2$. For further discussions of this assumption, see (8; 13).

1 The explicit expression for $P(\mathbf{S}\backslash S_i)$ is

$$P(\mathbf{S}\backslash S_i) = \frac{\prod_{j\neq i} \rho(S_j) \exp\left(\sum_{j>k;j,k\neq i} S_j J_{jk} S_k + \sum_{k,k\neq i} \theta_k S_k\right)}{\int \prod_{j\neq i} [dS_j \rho(S_j)] \exp\left(\sum_{j>k;j,k\neq i} S_j J_{jk} S_k + \sum_{k,k\neq i} \theta_k S_k\right)} \;.$$

Using the Gaussian assumption it is straightforward to derive the joint distribution of S_i and h_i

$$\begin{aligned} P(S_i, h_i) &= \frac{1}{Z_0^{(i)}} \rho(S_i)\, e^{S(h_i+\theta_i)}\, P(h_i \backslash S_i) \\ &= \frac{1}{Z_0^{(i)}} \frac{\rho(S_i)}{\sqrt{2\pi V_i}} \exp\left(S(h_i+\theta_i) - \frac{(h_i - \langle h_i \rangle_{\backslash i})^2}{2V_i} \right) \ , \end{aligned} \tag{1.7}$$

where the single variable partition function for S_i is defined as

$$\begin{aligned} Z_0^{(i)} &= \int dh_i dS_i\ \rho(S_i)\ e^{S(h_i+\theta_i)}\ P(h_i \backslash S_i) \\ &= \int dS\ \rho(S)\ e^{S(\langle h_i \rangle_{\backslash i} + \theta_i) + \frac{V_i}{2} S^2} \ . \end{aligned} \tag{1.8}$$

From (1.7) we immediately get the marginal distribution for S_i

$$P_i(S) = \frac{1}{Z_0^{(i)}} \rho(S_i)\ e^{S_i(\langle h_i \rangle_{\backslash i} + \theta_i) + \frac{V_i}{2} S_i^2} \ , \tag{1.9}$$

The expectation of S_i can now be written as

$$\langle S_i \rangle = \frac{\partial}{\partial \theta_i} \ln Z_0^{(i)} \qquad \text{for} \quad i = 1, \ldots, N \ . \tag{1.10}$$

(1.10) gives the first set of the TAP equations. The remaining task is to close the system of mean field equations by deriving expressions for $\langle h_i \rangle_{\backslash i}$ and $V_i = \langle h_i^2 \rangle_{\backslash i} - \langle h_i \rangle_{\backslash i}^2$. The first problem is easily solved by computing the mean of h_i as

$$\langle h_i \rangle = \int dh_i dS_i\ h_i\ P(S_i, h_i) = \langle h_i \rangle_{\backslash i} + V_i \langle S_i \rangle \ , \tag{1.11}$$

where the last equality is obtained from the explicit expression (1.7). Hence, we may eliminate $\langle h_i \rangle_{\backslash i}$ in favour of the mean field variables $\langle S_i \rangle$ and V_i, $i = 1, \ldots, N$ via

$$\langle h_i \rangle_{\backslash i} = \sum_j J_{ij} \langle S_j \rangle - V_i \langle S_i \rangle \ . \tag{1.12}$$

The last term in (1.12) is usually called the *Onsager reaction* term.

So far, all arguments are fairly standard. The main new contribution of this chapter is our method for computing the variances V_i.

The adaptive Onsager correction

The cavity variance of the field h_i is by definition given by

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

A naive application of our basic assumption, i.e. the weak dependencies of the variables S_j would lead us to neglecting the non-diagonal terms $j \neq k$. If we also neglect the fact that we are performing a cavity average, we arrive at

$$V_i \approx \sum_j J_{ij}^2 \left(\langle S_j^2 \rangle - \langle S_j \rangle^2 \right)_{\setminus i} \approx \sum_j J_{ij}^2 \left(\langle S_j^2 \rangle - \langle S_j \rangle^2 \right) \ . \tag{1.13}$$

This gives in fact the correct TAP equations for the SK-model (21). However, it is not expected to be correct for other models. Take e.g. a simple Gaussian model with $\rho_i(S) \propto e^{-\frac{1}{2}S_i^2}$ for which all expectations can be calculated analytically, and for which also our main TAP assumption of a Gaussian cavity distribution (1.6) is trivially fulfilled. If the J_{ij}'s are chosen as zero mean independent random variables with variance $\mathcal{O}(1/N)$ for $i < j$, again (1.13) is the right answer, in the sense that it gives the correct statistical physics for $N \to \infty$. This is no longer true if we introduce couplings which have weak higher order correlations. E.g. we may define the couplings by $J_{ij} = \frac{\beta}{N} \sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu$ as for the Hopfield model. Now the ξ_i^μ are iid random variables with zero mean and unit variance. In this case the covariance still fulfills the condition

$$\lim_{N \to \infty} \frac{1}{N^2} \sum_{ij} \left(\langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle \right)^2 = 0 \tag{1.14}$$

of weak correlations, but (1.13) is the wrong result.

We will next give a recipe how to compute the proper V_i's adaptively, i.e. without explicit knowledge of the distribution of the J_{ij}'s. Our method was motivated by work of (15) who have developed TAP equations for non iid statistics of the J_{ij}'s. We expect that our method will yield the correct statistical mechanics for fully connected models in the limit $N \to \infty$ for a large class of random matrix ensembles for the J_{ij}'s. We begin by defining the covariance matrix

$$\chi_{ij} \equiv \langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle = \frac{\partial \langle S_i \rangle}{\partial \theta_j} \tag{1.15}$$

where the second equality follows by direct differentiation with respect to θ_j. We develop a self-consistent computation of (1.15) based on (1.10) and (1.12).

We make one further approximation in the following: By differentiating the mean field equations we will keep the variances V_i fixed. We get

$$\begin{aligned}
\chi_{ij} &= \frac{\partial \langle S_i \rangle}{\partial \theta_j} \\
&= \frac{\partial \langle S_i \rangle}{\partial \theta_i} \frac{\partial \theta_i}{\partial \theta_j} + \frac{\partial \langle S_i \rangle}{\partial \langle h_i \rangle_{\setminus i}} \frac{\partial \langle h_i \rangle_{\setminus i}}{\partial \theta_j} \\
&= \frac{\partial \langle S_i \rangle}{\partial \theta_i} \left(\delta_{ij} + \frac{\partial \langle h_i \rangle_{\setminus i}}{\partial \theta_j} \right) \;, \qquad (1.16)
\end{aligned}$$

where the first line follows from differentiating eq. (1.10). Further using eq. (1.12), we finally get

$$\chi_{ij} = \frac{\partial \langle S_i \rangle}{\partial \theta_i} \left(\delta_{ij} + \sum_k (J_{ik} - V_k \delta_{ik}) \chi_{kj} \right) \qquad (1.17)$$

which is a linear equation for the matrix χ and can be solved with the result

$$\chi_{ij} = \left[(\mathbf{\Lambda} - \mathbf{J})^{-1} \right]_{ij} \;, \qquad (1.18)$$

where

$$\mathbf{\Lambda} \equiv \mathrm{diag} \left(V_1 + \left(\frac{\partial \langle S_1 \rangle}{\partial \theta_1} \right)^{-1}, \ldots, V_N + \left(\frac{\partial \langle S_N \rangle}{\partial \theta_N} \right)^{-1} \right) \;. \qquad (1.19)$$

The diagonal elements $\chi_{ii} = \langle S_i^2 \rangle - \langle S_i \rangle^2$ can also be obtained from

$$\chi_{ii} = \frac{\partial^2}{\partial \theta_i^2} \ln Z_0^{(i)} = \frac{\partial \langle S_i \rangle}{\partial \theta_i} \;. \qquad (1.20)$$

Demanding self-consistency, we obtain the additional set of TAP equations for the V_i's

$$\frac{\partial^2}{\partial \theta_i^2} \ln Z_0^{(i)} = \left[(\mathbf{\Lambda} - \mathbf{J})^{-1} \right]_{ii} \qquad \text{for} \quad i = 1, \ldots, N \qquad (1.21)$$

with $\mathbf{\Lambda}$ given by eq. (1.19). Note, that this result also holds exactly for a Gaussian model. Eqs. (1.10) for the $\langle S_i \rangle$'s have a computational complexity of $\mathcal{O}(N^2)$ whereas the complexity of computing the V_i's from (1.21) is $\mathcal{O}(N^3)$.

It is interesting to note that the functional form of the Onsager correction (1.21) expressed in terms of the $\langle S_i \rangle$'s and $\langle S_i^2 \rangle$'s (via χ_{ii}) is independent of the single variable density ρ_i. This supports a corresponding assumption made in (15) who derived TAP equations for an Ising model with a specific distribution of J_{ij}'s by computing the Onsager term (within a Gibbs free energy expansion) for a solvable model with spherical constraint.

Comparison with the variational mean field approach

In the variational MF approach, the distribution $P(\mathbf{S})$ is approximated by the factorized distribution $Q(\mathbf{S}) = \prod_i Q_i(S_i)$ which minimizes the *Kullback-Leibler* divergence

$$KL = \int d\mathbf{S}\; Q(\mathbf{S}) \ln \frac{Q(\mathbf{S})}{P(\mathbf{S})} = \ln Z + E[Q] - S[Q]\; . \tag{1.22}$$

For the model (1.1) we have

$$S[Q] = -\sum_i \int dS_i\; Q_i(S_i) \ln \frac{Q_i(S_i)}{\rho(S_i)} \tag{1.23}$$

and

$$E[Q] = -\sum_{i<j} J_{ij} m_i m_j - \sum_i m_i \theta_i \tag{1.24}$$

with $m_i = \langle S_i \rangle_Q$ and the bracket denotes expectation with respect to the distribution Q. One can easily perform the full minimization of (1.22) with the result

$$Q_i(S_i) = \frac{\rho(S_i)}{Z_0^{(i)}}\; e^{S_i(\sum_j J_{ij} m_j + \theta_i)}\;\; , \tag{1.25}$$

where

$$Z_0^{(i)} = \int dS_i\; \rho(S_i) e^{S_i(\sum_j J_{ij} m_j + \theta_i)} \tag{1.26}$$

A comparison with eqs. (1.9), (1.10) and (1.12) shows that this simple or 'naive' (as we will call it in the following) mean field theory is recovered from the TAP equations by setting the Onsager term

$$V_i = 0 \ . \tag{1.27}$$

1.4 Solving the TAP equations

In table 1.1, we give a recipe for solving the mean field equations by iteration. To simplify the notation we introduce the functions for computing the first moments from eq. (1.10) and the second moments from eq. (1.20)

$$f(\langle h_i \rangle_{\setminus i}, V_i) \;=\; \frac{\partial}{\partial \theta_i} \ln Z_0^{(i)} \quad (\; = \langle S_i \rangle \;) \tag{1.28}$$

$$f'(\langle h_i \rangle_{\setminus i}, V_i) \;=\; \frac{\partial \langle S_i \rangle}{\partial \theta_i} \quad (\; = \chi_{ii} \;) \ . \tag{1.29}$$

It turns out that the solution for the V_i's, are quite insensitive to the

Table 1.1
Pseudo-code for the general TAP mean field algorithm.

Initialization:
- Start from *tabula rasa*: $\langle \mathbf{S} \rangle := 0$ (or small random values if $\langle \mathbf{S} \rangle = 0$ is a fixed point).
- $\mathbf{V} := 0$ (ensemble learning estimate).
- Learning rate, $\eta := 0.05$.
- Fault tolerance, ftol $:= 10^{-3}$.
- variance_update = 20.

Iterate:

do:

for all i**:** (Equations (1.10) and 1.12))

$$\langle h_i \rangle_{\setminus i} := \sum_j J_{ij} \langle S_j \rangle - V_i \langle S_i \rangle$$

$$\delta \langle S_i \rangle := f(\langle h_i \rangle_{\setminus i}, V_i) - \langle S_i \rangle$$

endfor

for all i**:**

$$\langle S_i \rangle := \langle S_i \rangle + \eta \delta \langle S_i \rangle$$

for every variance_update iteration**:**

for all i**:** (Equation (1.19))

$$\Lambda_i := V_i + \frac{1}{f'(\langle h_i \rangle_{\setminus i}, V_i)}$$

for all i**:** (Equation (1.21))

$$V_i := \frac{1}{\left[(\mathbf{\Lambda} - \mathbf{J})^{-1} \right]_{ii}} - \Lambda_i$$

while $\max_i |\delta \langle S_i \rangle|^2 >$ ftol specific values of the mean spins $\langle S_i \rangle$. One may therefore update V_i greedily and more seldom, thus avoiding to perform the most computationally expensive operation so often. An important contributing factor to ensure fast convergence is the use of an adaptive learning rate: We set $\eta := 1.1\eta$ if 'the error' $\sum_i |\delta\langle S_i \rangle|^2$ decreases in the update step and $\eta := \eta/2$ otherwise.

The 'naive' mean field equations are obtained by omitting the V_i updating step. Estimates for the covariances, i.e. linear response corrections (14; 6; 4) are obtained by calculating $\chi_{ij} = \left[(\mathbf{\Lambda} - \mathbf{J})^{-1}\right]_{ij}$ after convergence.

It is a common experience that the naive mean field method is more robust than the TAP approach, i.e. for certain set-ups the TAP equations fail to converge where iterating the naive equations readily finds a solution (20). This usually happens when we are in a complex 'spin glass' like phase of the model (10). Discussions of this subtle matters can be found in (8). For many problems where the model is well matched to the data, we do not expect such a complex behaviour and the TAP equations are only slightly more difficult to solve than naive mean field equations (13). For a toy problem with replica symmetry (8) we found that the TAP equations reproduce the theoretical predictions to high accuracy (11).

1.5 Example I: The Hopfield model

The variables of the Hopfield model are Ising spins, i.e. $\rho(S_i) = \frac{1}{2}\delta(S_i - 1) + \frac{1}{2}\delta(S_i + 1)$. We thus get Ising model mean field equations (valid also for the SK-model and the Boltzmann machine)

$$\langle S_i \rangle = \tanh\left(\sum_j J_{ij}\langle S_j \rangle - V_i \langle S_i \rangle\right) \tag{1.30}$$

and $\chi_{ii} = 1 - \langle S_i \rangle^2$. In the Hopfield case the coupling matrix is given by $J_{ij} = \frac{\beta}{N}\sum_{\mu=1}^{\alpha N} \xi_i^\mu \xi_j^\mu$. We will breifly discuss how to obtain a simplified expression for the V_is in the case where the ξ_i^μs are iid random variables with zero mean and unit variance. We use the fact that the covariance (1.18) can be represented as an average over an auxiliary Gaussian measure via

$$\chi_{ii} = \left[(\mathbf{\Lambda} - \mathbf{J})^{-1}\right]_{ii} = -2\frac{\partial}{\partial \Lambda_i} \ln\left\{\int \frac{\prod_{i=1}^N dz_i}{(2\pi)^{N/2}} e^{-\frac{1}{2}\mathbf{z}^T(\mathbf{\Lambda}-\mathbf{J})\mathbf{z}}\right\} . \quad (1.31)$$

In the limit $N \to \infty$, we can assume that the V_i's do not fluctuate and can be replaced by a constant value value V which can be computed by averaging (1.31) over the distribution of the ξ_i^μ's.

A straightforward calculation (the details of which will be presented elsewhere) shows that

$$V = \frac{\beta\alpha}{1 - \beta(1-q)} \quad (1.32)$$

with $q = \frac{1}{N}\sum_i \langle S_i \rangle^2$ in accordance with the result of chapter XIII in (8) and also derived in chapter **Kabashima+Saad**.

1.6 Example II: Bayesian learning with a perceptron

In this section, we try to be a bit more ambitious. Our derivation of the TAP equations was designed for computing expectations with proper probability measures. However, it is tempting to apply the TAP equations to integrals over *complex functions* which commonly arise when we transform complicated probabilistic models into simpler ones involving auxiliary integration variables. As an example, we discuss the TAP approach to classification with a simple Bayesian model. More details about the validity of this approach will be given elsewhere.

The basic set-up of the Bayesian learning model is as follows: We have a training set

$$\mathcal{D}_N = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, N\} \quad (1.33)$$

of input vectors $\mathbf{x}_i$ and associated output labels $y_i \in \{-1, +1\}$. We assume that the likelihood of the outputs y conditioned on input $\mathbf{x}$ is parametrized by a d dimensional weight vector $\mathbf{w}$ and we use a probit model[2]

$$P(y|\mathbf{x}) = \Phi\left(y\frac{\mathbf{w}\cdot\mathbf{x}}{\sigma}\right) . \quad (1.34)$$

2 This choice corresponds to the likelihood $p(y|S) = \Theta(y(S+\xi))$ with $\Theta(x) = 1$ for $x > 0$ and 0 otherwise and ξ being Gaussian noise with variance σ^2.

Φ is the Gaussian cumulative distribution function

$$\Phi(z) = \int_{-\infty}^{z} \frac{dt}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \ . \tag{1.35}$$

A Gaussian prior distribution over the weights

$$P(\mathbf{w}) = (2\pi)^{-N/2} e^{-\frac{1}{2}\mathbf{w}\cdot\mathbf{w}} \tag{1.36}$$

leads for the set (1.33) to the posterior density

$$P(\mathbf{w}|D_N) \propto e^{-\frac{1}{2}\mathbf{w}\cdot\mathbf{w}} \prod_{i=1}^{N} \Phi\left(y_i \frac{\mathbf{w}\cdot\mathbf{x}_i}{\sigma}\right) \ . \tag{1.37}$$

Our goal is to compute the expected posterior weight vector $\langle \mathbf{w} \rangle$ and use sign$\langle \mathbf{w} \rangle \cdot \mathbf{x}$ as an approximation to the Bayes classifier (11). Unfortunately, (1.37) does not have the form of the distribution (1.1). By introducing an integral representation for the Dirac δ-function

$$\Phi(z) = \int dh \Phi(h) \int_{-i\infty}^{+i\infty} \frac{dS}{2\pi} e^{S(z-h)} \ , \tag{1.38}$$

we can integrate over $\mathbf{w}$. An integration by parts yields the representation

$$\langle \mathbf{w} \rangle = \sum_i \alpha_i \mathbf{x}_i \ , \tag{1.39}$$

with

$$\alpha_i = \langle S_i \rangle = \frac{1}{Z} \int \prod_j (dS_j \rho(S_j)) \ S_i \ \exp\left[\frac{1}{2}\sum_{ij} S_i J_{ij} S_j + \sum_i \theta_i S_i\right] . \tag{1.40}$$

The couplings are defined as

$$J_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j \tag{1.41}$$

for $i \neq j$ and $J_{ii} = 0$. The single variable distribution is given by

$$\rho(S_i) = \frac{1}{2\pi} e^{\frac{1}{2}\mathbf{x}_i\cdot\mathbf{x}_i S_i^2} \int dh \ \Phi\left(\frac{y_i h}{\sigma}\right) \ e^{-hS} \ . \tag{1.42}$$

We have also introduced extra external fields θ_i which have to be set equal to zero at the end of the calculations.

We are ready to get the TAP equations from (1.10) and (1.21) where

$$Z_0^{(i)} = \int_{-i\infty}^{+i\infty} dS\ \rho(S)\ e^{S(\langle h_i\rangle_{\setminus i}+\theta_i)+\frac{V_i}{2}S_i^2} = \Phi\left(y_i \frac{\langle h_i\rangle_{\setminus i}+\theta_i}{\sqrt{V_i+\mathbf{x}_i\cdot\mathbf{x}_i+\sigma^2)}}\right).$$

It would be nice if we could demonstrate how well our method approximates the true averages for a real data problem. So far, we have not performed the extensive Monte Carlo simulations needed for such an experiment.

It is possible to present a weaker result by showing that our approach is at least internally consistent. We display the cavity averages $y_i\langle h_i\rangle_{\setminus i}$ for the 'Sonar' dataset (3) in Figure 1.1. These have been computed in two different ways. First, we use their literal definitions and leave S_i (corresponding to the i-th data point) out of the system. We then compute the expected field $\langle h_i\rangle_{\setminus i}^{\text{exact}}$ for the $N-1$ variable system by solving the TAP equations on this reduced set. On the other hand, from eq. (1.12) we get, $\langle h_i\rangle_{\setminus i}$, the mean field estimate of of the cavity mean. Ideally, both calculations should give the same result and when plotted against each other (as in figure 1.1), the points should lie on the diagonal. We have compared our adaptive TAP approach (squares) with an approach which uses a constant value for $V_i = V$ (triangles) based on the assumption that all $\mathbf{x}_i$ are independently drawn from a spherical Gaussian. Our adaptive algorithm is superior to the simpler method especially at negative values of $y_i\langle h_i\rangle_{\setminus i}$. This is important, because the cavity field can be used to compute a leave-one-out estimate of the test error for our Bayesian classifier by counting the number of examples where $y_i\langle h_i\rangle_{\setminus i} < 0$ (11; 13). Figure 1.2 shows the distribution of V_i's obtained from the adaptive algorithm. For comparison we show the constant value V found in the non-adaptive method as a vertical thick line.

Figure 1.3 displays test error rates for the 'naive' mean field algorithm (which was computed by setting the Onsager term $V_i = 0$) and the TAP MF method for the Bayesian classifier. The data are obtained for a toy problem, where the inputs are drawn independently from a spherical Gaussian (with zero mean and variance $1/d$) and the labels are generated by a 'teacher' perceptron (11). The differences between the two mean field theories are small in this case. Note, that a proper variational mean field method based on using the Gaussian process for-

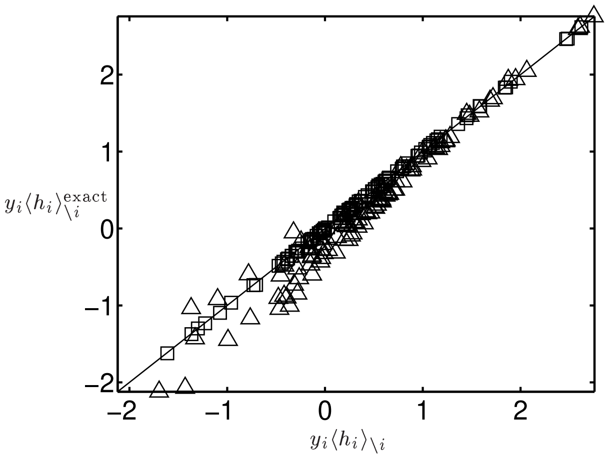

Figure 1.1
Comparison between the exact mean cavity field and the one computed from the TAP equations for the sonar data. Squares are for the adaptive theory and triangles for constant V theory.

mulation of (2) is not possible for this model for $N > d$ (and $\sigma = 0$), because the kernel matrix is singular.

1.7 Outlook

We have derived an adaptive TAP MF approximation for models with quadratic interactions. It is based on the assumption of a large number of weakly interacting random variables (this may in practice be achieved only, when the effect of a few 'strong' variables is separated **Explain a bit more here**), but does not assume a specific distribution for the couplings.

At present we are working on generalizations of our approach in various directions:

• It is possible to map a variety of complicated probabilistic models onto models with quadratic interactions involving auxiliary variables. In general, the models in the augmented set of variables will involve complex functions and the cavity derivation must be adapted to this new framework.

• We will further show that our MF theory becomes exact for large classes of random matrix ensembles for couplings in the 'thermodynamic' limit of large systems.

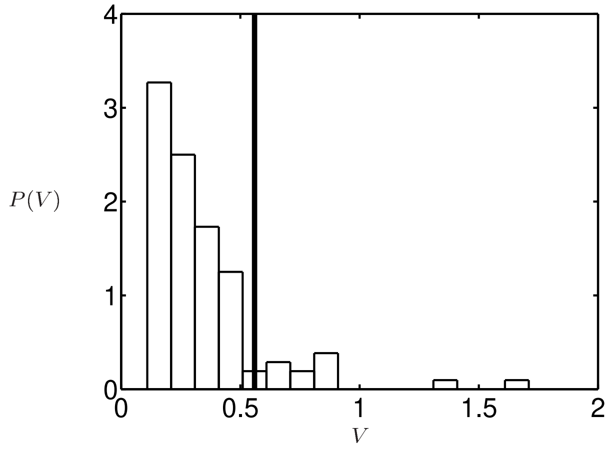

Figure 1.2
The distribution of V_i for the adaptive theory for the sonar data. The thick vertical bar is the value found by the constant V theory.

• We will develop a TAP approximation for the Gibbs free energy of the quadratic models. This is not only of practical relevance for model selection, but will also provide a derivation of our adaptive TAP method from a variational principle. This may help to improve algorithms for solving the TAP equations.

References

[1]Ackley, D. Hinton, G., and Sejnowski, T., A Learning Algorithm for Boltzmann Machines, Cognitive Science **9**, 147-169 (1985).

[2]Csató, L., Fokoué, E., Opper, M., Schottky, B., and Winther, O., Efficient Approaches to Gaussian Process Classification, in Advances in Neural Information Processing Systems 12 (NIPS'99), Eds. Solla S. A., Leen T. K., and Mü ller K.-R., MIT Press (2000).

[3]Gorman, R. P. and Sejnowski T. J., Neural Networks **1**, 75 (1988).

[4]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).

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

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

[7]Lee, T.-W., *Independent Component Analysis*, Kluwer Academic Publishers, Boston (1998).

[8]Mézard, M. , Parisi, G., and Virasoro, M. A., *Spin Glass Theory and Beyond*, Lecture Notes in Physics, 9, World Scientific (1987).

[9]Mézard, M. and Parisi, G., Europhys. Lett. **2**, 913 (1986).

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

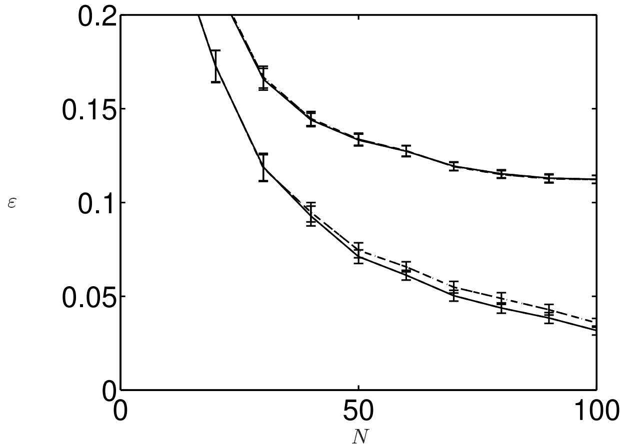

Figure 1.3
Comparison of TAP and naive mean field theory. Learning curves – test set error rates ε versus the number of examples N with $d = 10$. The upper curves (on top of each other) are for the noisy Bayesian scenario $\sigma^2 = 0.01$ (noisy teacher). The lower curves are for the noise-free scenario. The dashed line are for naive mean field theory.

[11]Opper, M., and Winther, O., A Mean Field Approach to Bayes Learning in Feed-Forward Neural Networks, Phys. Rev. Lett. **76**, 1964 (1996).

[12]Opper, M., and Winther O., A mean field algorithm for Bayes learning in large feed-forward neural networks, Advances in Neural Information Processing Systems 9 (NIPS'96), Eds. M. C. Mozer, M. I. Jordan, and T. Petsche, 225-331, MIT Press (1997).

[13]Opper, M., and Winther, O., Gaussian Processes for Classification: Mean Field Algorithms, Neural Computation (2000).

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

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

[16]Peterson, C., and Anderson, J., A Mean Field Learning Algorithm for Neural Networks, Complex Systems **1**,995-1019 (1987).

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

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

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

[20]Tanaka, T., Mean-Field Theory of Boltzmann Machine Learning, Phys. Rev. E **58**, 2302-2310 (1998).

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