The following full text is a preprint version which may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/112690

Please be advised that this information was generated on 2018-01-14 and may be subject to change.
1 Mean field theory for graphical models

Hilbert J. Kappen and Wim J. Wiegerinck
RWCP Theoretical Foundation SNN University of Nijmegen

In this chapter, mean field theory is introduced from an information theoretic viewpoint. The mean field approximation is defined as the factorized distribution that is closest to the target distribution. When using the KL divergence to define closeness, this factorized distribution must have equal marginals as the target distribution. Such marginals can be approximately computed by using a Taylor series expansion in the couplings around the factorized distribution. To lowest order in the couplings, the usual naive mean field equations are obtained and to second order, one obtains the TAP equations. An important advantage of this procedure is that it does not require the concept of a free energy. Therefore, it can be applied to arbitrary probability distributions, such as arising in asymmetric stochastic neural networks and graphical models.

1.1 Introduction

During the last few years, the use of probabilistic methods in artificial intelligence and machine learning has gained enormous popularity. In particular, probabilistic graphical models have become the preferred method for knowledge representation and reasoning [1]. The advantage of the probabilistic approach is that all assumptions are made explicit in the modeling process and that consequences, such as predictions on novel data, are assumption free and follow from a mechanistic computation. The drawback of the probabilistic approach is that the method is intractable. This means that the typical computation scales exponentially with the problem size.

Recently, a number of authors have proposed methods for approximate inference in large graphical models. The simplest approach gives a lower bound on the probability of a subset of variables using Jenssen's inequality [2]. The method involves the minimization of the KL divergence between the target probability distribution \( p \) and some 'simple' variational distribution \( q \). The method can be applied to any probability model, whether directed or undirected.
The Boltzmann-Gibbs distributions is widely used in physics, and mean field theory has been known for these distributions for a long time. For instance, for the Ising model on a square lattice, it is known as the Bragg-Williams approximation [3] and it is generalized to other models in the Landau theory [4]. One can show that the above lower bound corresponds to the first term in a Taylor series expansion of the free energy around a factorized model. This Taylor series can be continued and the second order term is known as the Thouless Anderson Palmer (TAP) correction [5, 6, 7, 8]. The second order term significantly improves the quality of the approximation, depending on the amount of frustration in the system, but is no longer a bound.

For probability distributions that are not Boltzmann-Gibbs distributions, it is not obvious how to obtain the second order approximation. However, there is an alternative way to compute the higher order corrections, based on an information theoretic argument. The general approach to this mean field approximation is introduced in section 1.2. Before, we work out the mean field approximations for the general case, we first illustrate this idea for Boltzmann distributions in section 1.3. Subsequently, in section 1.4 we consider the general case. Finally, in section 1.5 we illustrate the approach for sigmoid belief networks.

1.2 Mean field theory

In this section we consider a form of mean field theory that was previously proposed by Plefka [6] for Boltzmann-Gibbs distributions. It turns out, however, that the restriction to Boltzmann-Gibbs distributions is not necessary and one can derive results that are valid for arbitrary probability distributions. We therefore consider the general case.

Our argument uses an information geometric viewpoint. For an introduction to this approach see for instance [9]. Let $x = (x_1, \ldots, x_n)$ be an $n$-dimensional vector, with $x_i$ taking on discrete values. Let $p(x; \Theta)$ be a probability distribution on $x$, parametrized by $\Theta$. Let $\mathcal{P} = \{p(x; \Theta)\}$ be the manifold of all the probability distributions that can be obtained by considering different values of $\Theta$.

We now assume that that $\mathcal{P}$ contains a submanifold of factorized probability distributions in the following sense. We assume that the parametrization $\Theta$ is such that it can be divided into two subsets,
Mean field theory for graphical models

\[ \Theta = (\theta, w), \] and that the submanifold \( \mathcal{M} \subset \mathcal{P} \) of factorized probability distributions is described by \( w = 0 \). \( \theta \) parametrizes the factorized distributions in the manifold \( \mathcal{M} \), and \( w \) parametrizes the remainder of the manifold \( \mathcal{P} \). We will denote factorized distributions by \( q(x|\theta) = p(x|\theta, w = 0) \).

Consider an arbitrary probability distribution \( p(x|\theta, w) \in \mathcal{P} \). We define its mean field approximation as the factorized distribution \( q(x|\theta^q) \in \mathcal{M} \) that is closest to \( p(x|\theta, w) \). As a distance measure, we use the Kulback-Leibler divergence \( [9, 10] \)

\[
KL = \sum_x p(x|\theta, w) \log \frac{p(x|\theta, w)}{q(x|\theta^q)}. \tag{1.1}
\]

Since \( q(x|\theta^q) \) is a factorized distribution, \( q(x|\theta^q) = \prod_{i=1}^n q_i(x_i|\theta_i^q) \), we can find the closest \( q \) by differentiating the Kulback-Leibler divergence with respect to these independent components \( q_i(x_i|\theta_i^q) \). Using a Lagrange multiplier to ensure normalisation of \( q_i(x_i|\theta_i^q) \), one finds that this optimal \( q \) must satisfy

\[
q_i(x_i|\theta_i^q) = p(x_i|\theta, w), \tag{1.2}
\]

where \( p(x_i|\theta, w) \) is the marginal distribution of \( p(x|\theta, w) \) on variable \( x_i \).

Thus, we are looking for a factorized distribution that has the same marginals over single variables as the target distribution \( p \). However, since \( p \) is intractable, we cannot compute its marginals. Therefore, Eq. 1.2 as it stands is not very helpful.

Assume now that \( p(x_i|\theta, w) \) is somehow close to the factorized submanifold. We can expand \( p(x_i|\theta, w) \) around \( q_i(x_i|\theta_i^q) \) in terms of changes in the parameters \( \Delta \theta = \theta - \theta^q \) and \( \Delta w = w - 0 \). We define \( \Delta \Theta = (\Delta \theta, \Delta w) \). Instead of expanding \( p(x_i) \), we prefer to expand \( \log p(x_i) \) as it will turn out to simplify the derivations. Thus,

\[
\log p(x_i|\theta, w) = \log q_i(x_i|\theta_i^q) + \Delta \log p(x_i),
\]

with \( \Delta \log p(x_i) = \sum_j \left( \frac{\partial \log p(x_i)}{\partial \Theta_j} \right)_q \Delta \Theta_j \).

\[ \footnote{Note, that to obtain the standard variational bound using Jensen's inequality, one employs "the other" KL divergence with the roles of \( p \) and \( q \) reversed. As will be outlined below, the KL divergence considered here gives the same result as the Jensen's bound to lowest order.} \]
\[ + \frac{1}{2} \sum_{JK} \left( \frac{\partial^2 \log p(x_i)}{\partial \Theta_J \partial \Theta_K} \right)_q \Delta \Theta_J \Delta \Theta_K \]
+ higher order terms

(1.3)

The differentials are evaluated at the factorized distribution \( q \). They can be evaluated in polynomial time for the graphical models that we discuss in the following sections. The left hand side of Eq. 1.3 is zero, because of Eq. 1.2. We therefore can solve Eq. 1.3 in terms of the unknowns \( \theta_i^q \). The resulting factorized distribution \( q(x_i | \theta_i^q) \) gives the desired marginals up to the order of the expansion of \( \Delta \log p(x_i) \).

1.3 Boltzmann distributions

As an example, consider the manifold of Boltzmann distributions

\[ p(x|w, \theta) = \frac{1}{Z} \exp \left( \frac{1}{2} \sum_{ij} w_{ij} x_i x_j + \sum_i \theta_i x_i \right). \]

with \( x_i = \pm 1 \). \( Z \) is a normalization constant,

\[ Z = \sum_s \exp(-E(s)) \]

(1.5)

and is called the partition function.

For any \((\theta, w)\), we can in principle compute the statistics such as the mean values \( m_i = \langle x_i \rangle = \sum_x x_i p(x|\theta, w) \) and correlations \( \chi_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle \). Both \((\theta, w)\) and \((m, \chi)\) form a coordinate system of \( P \), which means that any \( p \in P \) can be specified by either \((\theta, w)\) or \((m, \chi)\).

The situation is illustrated in Fig. 1.1. We consider the manifold of Boltzmann distribution on two variables \( x_1 \) and \( x_2 \), parametrized by \( \theta_1 = \theta_2 = \theta \) and \( w_{12} = w_{21} = w \). Every distribution is specified by a point in the \( \theta, w \) plane. At each point \((\theta, w)\), we can compute \( \langle x_1 \rangle = \langle x_2 \rangle = m \) (upper left) as well as the correlation \( \langle x_1 x_2 \rangle - \langle x_1 \rangle \langle x_2 \rangle = \chi \) (upper right). The lines of constant \( m \) and \( \chi \) are superimposed in the bottom figure. Any \( p \in P \) is equally well specified by its 'couplings' \((\theta^p, w^p)\) or by its 'statistics' \((m^p, \chi^p)\). The submanifold of factorized distributions is given by \( w = 0 \). The mean field approximation for \( p \) is found by the intersection of the line \( w = 0 \) and the constant \( m^p \) contour.

Since \( x_i \) is binary, the marginals on \( x_i \) are completely determined in
Mean field theory for graphical models

Figure 1.1
Statistics \( m \) (upper left) and \( \chi \) (upper right) as a function of \( \theta \) and \( w \) for a Boltzmann distribution on two variables \( (x_1, x_2) \), with \( w_{12} = w_{21} = w \) and \( \theta_1 = \theta_2 = \theta \). Below: Contour lines of constant \( m \) (solid) and constant \( \chi \) (dashed).

Both \( (\theta, w) \) and \( (m, \chi) \) are coordinates systems of \( \mathcal{P} \). \( \mathcal{M} \) is given by the line \( w = 0 \).

For any \( p \in \mathcal{P} \), the closest \( q \in \mathcal{M} \) satisfies \( m^q = m^p \).

terms of the expected value of \( x_i \):

\[
p(x_i|\theta, w) = \frac{1}{2}(1 + m^q_i) x_i \quad (1.6)
\]

\[
q_i(x_i|\theta^q) = \frac{1}{2}(1 + m^q_i) x_i. \quad (1.7)
\]

\( m^p,q \) are the expectation values of \( x_i \) under the distributions \( p \) and \( q \), respectively. Therefore, Eq. 1.2 becomes

\[
m_i^q = m_i^p. \quad (1.8)
\]

Since \( q \) is a factorized distribution, \( m_i^q = \tanh \theta_i^q \). Since \( p \) is an intractable distribution, no such simple relation exists for \( m_i^p \) in terms of \( \theta \) and \( w \). Eq. 1.3 becomes

\[
0 = \Delta m_i \approx \sum_j \frac{\partial m_i}{\partial \theta_j} |q \Delta \theta_j + \frac{1}{2} \sum_{j,K} \frac{\partial^2 m_i}{\partial \theta_j \partial \theta_K} |q \Delta \theta_j \Delta \theta_K, \quad (1.9)
\]

where \( \Theta_I = (\theta_i, w_{ij}) \) is the vector of all weights and thresholds.

For the Boltzmann distribution, it is easy to show that

\[
\langle x_i \rangle = \langle \tanh(h_i(x)) \rangle, \quad (1.10)
\]

with \( h_i = \sum_j w_{ij} x_j + \theta_i \). This equation allows us to compute the derivatives in the factorized point \( q \). This computation is tedious but straightforward. It is presented in appendix A. The result is

\[
m_i = \tanh(\sum_j w_{ij} m_j + \theta_i - m_i \sum_j w_{ij}^2 (1 - m^2_j)). \quad (1.11)
\]
Eq. 1.11 gives the approximate mean firing rates in terms of the couplings $\theta_i, w_{ij}$. It was first derived by Thouless, Anderson and Palmer and is referred to as the TAP equation [5]. It can also be obtained from a Taylor series expansion of the free energy at constant $m$ [6]. The motivation for expansion at constant $m$ follows naturally from our geometric argument. It was applied to Boltzmann machine learning in [7, 8].

It is interesting to note that Eq. 1.11 is not only true for Boltzmann distributions, but also for stochastic neural networks with asymmetric connectivity $w_{ij} \neq w_{ji}$. In the asymmetric case, the stationary distribution is not of the Boltzmann type. However, as long as the dynamics is ergodic, there still exists a unique, stationary distribution $p(x|\theta, w)$, but the analytic expression of $p(x|\theta, w)$ is not known. If the neural dynamics is sequential or parallel Glauber dynamics it is rather easy to show that Eq. 1.10, however, is still true, where the expectation value is with respect to the unknown stationary distribution $p(x|\theta, w)$. The differentials in Eq. 1.9 are computed from Eq. 1.10 and the result is identical as in the symmetric case [11].

### 1.4 The general case

In this section, we consider the problem of computing the mean field approximation for a probability distribution in the presence of evidence. Let $x = (x_1, \ldots, x_n)$ be an $n$-dimensional vector, with $x_i$ taking on discrete values. Let $p(x)$ be a graphical model on $x$. We will assume that $p(x)$ can be written as a product of potentials in the following way:

$$\log p(x|\phi) = \sum_{\alpha=1}^{m} \phi_{\alpha}(x^\alpha) - \Psi(\phi).$$

(1.12)

Here $x^\alpha$ denotes the subset of variables from $x_1, \ldots, x_n$ that appear in potential $\alpha$. Potentials can be overlapping, $x^\alpha \cap x^\beta \neq \emptyset$, and $x = \bigcup_{\alpha} x^\alpha$. $\Psi(\phi)$ is a constant, that ensures normalisation of $p(x|\phi)$. $m$ is the total number of potentials.

For directed graphical models, the potentials are given in terms of conditional probability distributions. Thus, there is one potential for each node and $\phi_i(x^i) = \log p(x_i|\pi_i)$, with $\pi_i$ the set of parents of node $i$. Since each potential ensures its own normalisation, $\Psi(\phi) = 0$. For Boltzmann distributions, the number of potentials is $m = n(n - 1)/2$. 
one for each coupling $w_{ij}$ and $\Psi$ is the log partition function.

We wish to compute the marginal probability that $x_i$ has some specific value $s_i$ in the presence of evidence. We therefore denote $x = (e, s)$ where $e$ denote the subset of evidence variables, and $s$ denotes the rest of the variables. The marginal is given as

$$p(s_i|e) = \frac{p(s_i, e)}{p(e)} \quad (1.13)$$

Both numerator and denominator contain sums over hidden states. These sums scale exponentially with the size of the problem, and therefore the computation of marginals is intractable.

We now think of the manifold of all probability distributions of the form Eq. 1.12, spanned by the coordinates $\phi = \{\phi_{\alpha}(x^\alpha), \alpha = 1, \ldots, m\}$. For each $\alpha$, $\phi_{\alpha}(x^\alpha)$ is a probability table, i.e. a table of numbers indexed by $x^\alpha$. Thus, $\phi$ can be thought of as a vector of probability tables. This manifold contains values of $\phi$ for which $p(x|\phi)$ is factorized. Let $\theta$ be such a value, then

$$\log p(x|\theta) = \sum \phi_{\alpha}(x^\alpha) - \Psi(\theta) = \sum_i \log q_i(x_i) = \log q(x). \quad (1.14)$$

$q(x)$ denotes this factorized distribution.

We can therefore use the mean field approach outlined in section 1.2 where the coordinates of the manifold $(\theta, w)$ are given by $\{\phi_{\alpha}(x^\alpha), \alpha = 1, \ldots, m\}$ and the factorized point $(\theta, 0)$ is given by $\log q(x)$. Therefore, we expand in $\Delta \phi_{\alpha}(x^\alpha) = \phi_{\alpha}(x^\alpha) - \log q(x^\alpha)$:

$$\Delta \log p(s_i|e) = \sum_{\alpha=1}^{m} \sum_{x^\alpha} \left( \frac{\partial \log p(s_i|e)}{\partial \phi_{\alpha}(x^\alpha)} \right) \Delta \phi_{\alpha}(x^\alpha)$$

$$+ \frac{1}{2} \sum_{\alpha, \beta=1}^{m} \sum_{x^\alpha, y^\beta} \left( \frac{\partial^2 \log p(s_i|e)}{\partial \phi_{\alpha}(x^\alpha) \partial \phi_{\beta}(y^\beta)} \right) \Delta \phi_{\alpha}(x^\alpha) \Delta \phi_{\beta}(y^\beta)$$

$$+ \text{higher order terms} \quad (1.15)$$

It is straightforward to compute the derivatives:

$$\frac{\partial \log p(s_i|e)}{\partial \phi_{\alpha}(x^\alpha)} = p(x^\alpha|s_i, e) - p(x^\alpha|e)$$

2 For instance, when $\alpha$ contains three binary variables, $\phi_{\alpha}(x^\alpha)$ is an 8-dimensional vector.
\[
\frac{\partial^2 \log p(s_i|e)}{\partial \phi_k(x^k) \partial \phi_l(\tilde{y}^l)} = p(x^k, \tilde{y}^l|s_i, e) - p(x^k|\tilde{y}^l|e) - p(x^k|s_i, e) p(\tilde{y}^l|s_i, e) + p(x^k|e) p(\tilde{y}^l|e)
\]

These derivatives must be evaluated at the factorized point \( q \). Denote \( \langle \cdot \rangle_{s_i,e} \) and \( \langle \cdot \rangle_e \) the expectation values with respect to the factorized distributions \( q(x|s_i, e) \) and \( q(x|e) \), respectively.

We first consider the first order mean field equation. Setting the lhs of Eq. 1.15 equal to zero, we obtain

\[
0 = \sum_k \langle \Delta \phi_k \rangle_{s_i,e} - \langle \Delta \phi_k \rangle_e = \sum_k \langle \phi_k \rangle_{s_i,e} - \log q_i(s_i) + \text{constant}
\]

The constant term collects all contributions that are independent of \( s_i \). These terms do not have to be considered, because they are absorbed in the normalisation of \( q_i \). Eq. 1.16 can be reformulated as

\[
q(s_i) = \frac{1}{Z_i} \exp \left( \langle \log p \rangle_{s_i,e} \right)
\]

in which the constants \( Z_i \) follow from normalisation. This expression is identical to the standard variational bound, obtained from the Jensen inequality.

The correction with second order terms is obtained as follows. From 1.16 we have

\[
\sum_k \langle \Delta \phi_k \rangle_{s_i,e} = \sum_k \langle \Delta \phi_k \rangle_e + O(\Delta^2)
\]

at the solution \( q \). Setting Eq. 1.15 to zero, combined with (1.18), we find the second order correction to the mean field equations

\[
q(s_i) = \frac{1}{Z_i} \exp \left( \langle \log p \rangle_{s_i,e} + \frac{1}{2} \sum_{k,l} \langle \Delta \phi_k \Delta \phi_l \rangle_{s_i,e} \right)
\]

These equations are generalization of the mean field equations with TAP corrections for directed graphical models.

**Complexity and single-overlap graphs**

The complexity of the first order mean field equations 1.17 is exponential in the number of variables in the potentials \( \phi_k \) of \( p \). For the approxi-
Mean field theory for graphical models

Figure 1.2
An example of a single-overlap graph. Left: The chest clinic model (ASIA)[12]. Right: nodes within one potential are grouped together, showing that potentials share at most one node.

formation of a directed graphical model, this means that the mean field method is only tractable if the number of parents is reasonably small. The complexity of the second order mean field equations 1.19 is exponential in the number of variables of the union of two overlapping potentials, due to the \((\Delta \phi_k \Delta \phi_l)_{s_1,e}\) term. This complexity scales at worst as the square of the complexity of the standard mean field approximation.

This is in contrast to Boltzmann machines, in which the TAP approximation has the same complexity as the standard mean field approximation. However, this result can be generalized to what we call single-overlap graphs, which are graphs in which the potentials \(\phi_k\) share at most one node. Figure 1.4 shows an example of a single-overlap graph.

For these graphs one can derive

\[
\log q(s_i) \propto \langle \log p \rangle_{s_1,e} + \frac{1}{2} \sum_{\alpha} \left( \langle (\Delta \phi_{\alpha})^2 \rangle_{s_1,e} - \langle \Delta \phi_{\alpha} \rangle_{s_1,e}^2 \right) + \sum_{\alpha, j \neq i} \left( \frac{1}{2} \langle (\Delta \phi_{\alpha})_{j_1,e} \rangle_{s_1,e} - \langle \Delta \phi_{\alpha} \rangle_{s_1,e} \right) \Delta \phi_{\alpha} \right)_{s_1,e} (1.20)
\]

which has a complexity that is of the same order as the complexity of the standard mean field approximation [13].

In any case, for large potentials, additional approximations are required, as was proposed by [2] for the first order mean field equations for sigmoid belief networks. It is evident, that such additional approximations are then also required for the second order mean field equations.
Table 1.1
Marginal probabilities of states being true obtained in the chest clinic model (ASIA). First column: exact marginals. Second column: marginals computed using first order approximation (mean field). Third column: marginals computed using all approximation up to second order (TAP).

<table>
<thead>
<tr>
<th>Node</th>
<th>Exact</th>
<th>MF</th>
<th>TAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visit to Asia?</td>
<td>0.010</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>Smoking?</td>
<td>0.500</td>
<td>0.420</td>
<td>0.524</td>
</tr>
<tr>
<td>Tuberculosis?</td>
<td>0.010</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Lung cancer?</td>
<td>0.050</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Bronchitis?</td>
<td>0.450</td>
<td>0.264</td>
<td>0.410</td>
</tr>
<tr>
<td>Either t or l?</td>
<td>0.065</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>positive X ray?</td>
<td>0.110</td>
<td>0.050</td>
<td>0.050</td>
</tr>
<tr>
<td>Dyspnoea?</td>
<td>0.428</td>
<td>0.223</td>
<td>0.407</td>
</tr>
</tbody>
</table>

Numerical results
We illustrate the theory by a toy problem, which is inference in Lauritzen's chest clinic model (ASIA), defined on 8 binary variables \{A, T, S, L, B, E, X, D\} (see figure 1.4, and [12] for more details about the model). We computed exact marginals and approximate marginals using the approximating methods up to first and second order respectively. The approximate marginals are determined by sequential iteration of (1.17) and (1.19), starting at \(q(x_i) = 0.5\) for all variables \(i\). Results are shown in table 1.4. We observe that the TAP approximation gives a significant improvement over the first order mean field result.

1.5 Sigmoid belief networks
In this section, we illustrate the geometric mean field approach for sigmoid belief networks on binary variables, defined as

\[
p(x) = \prod_i \sigma(x_i h_i),
\]

where \(\sigma(x) = (1 + \exp(-2x))^{-1}\), \(x_i = \pm 1\) and \(h_i\) is the local field: \(h_i(x) = \sum_{j=1}^n w_{ij} x_j + \theta_i\). Since nodes can have a large number of parents, computation of the marginal distributions for the sigmoid belief network is intractable in general. This is also the case in the presence of evidence.

When evidence is present, we separate the total set of variables into evidence variables \(e\) and the remaining hidden variables \(s\): \(x = (s, e)\).
Mean field theory for graphical models

The conditional distribution becomes

$$p(s|e) = \frac{p(s, e)}{p(e)} = \frac{\prod_{i \in e} \sigma(s_i h_i(s, e)) \prod_{i \in e} \sigma(e_i h_i(s, e))}{\sum_{s'} \prod_{i \in e} \sigma(s_i' h_i(s', e)) \prod_{i \in e} \sigma(e_i h_i(s', e))}.$$  \hspace{1cm} (1.22)

When couplings from hidden nodes to hidden or evidence nodes are zero, \( w_{ij} = 0 \), all \( i, j \in s \), the conditional probability reduces to a factorized distribution

$$p(s|e) \rightarrow q(s) = \prod_{i} \sigma(s_i \theta_i^q)$$  \hspace{1cm} (1.23)

In this case, the "evidence" factorizes as well:

$$p(e) \rightarrow r(e) = \prod_{i \in e} \sigma(e_i h_i \left( \sum_{j \in e} w_{ij} e_j + \theta_i \right))$$  \hspace{1cm} (1.24)

As outlined in section 1.4, we define the mean field approximation to \( p(s|e) \) as the factorized distribution \( q(s) \) that has identical marginals: \( p(s_i|e) = q(s_i) \). The solution is given in terms of the parameters \( \theta_i^q \) that specify \( q(s) \) such that \( \tanh(\theta_i^q) = m_i^q \). Since \( p(s|e) \) is intractable, we compute the marginals approximately by a Taylor series expansion around \( q \). Whereas in general for directed graphical models one must expand in terms of the conditional probability tables, one can expand sigmoid belief networks in terms of the couplings \( w_{ij} \), similar to the Boltzmann distribution.

The mean field equations are given by Eq. 1.3:

$$0 = \sum_{k, l} \frac{\partial \log p(s_i|e)}{\partial w_{kl}} \Delta w_{kl} + \frac{1}{2} \sum_{k, l, k', l'} \frac{\partial^2 \log p(s_i|e)}{\partial w_{kl} \partial w_{k'l'}} \Delta w_{kl} \Delta w_{k'l'}.$$  \hspace{1cm} (1.25)

The computation of the derivatives is straightforward, but tedious. The resulting mean field equations, including TAP corrections, are

\[
m_i = \tanh \left( \sum_{j \in s, e} m_{ij} w_{ij} + \theta_i + 2 \sum_{k \in e} r(-e_k) e_k w_{ki} \right)
\]

\[
-m_i \left( \sum_{k \in s} (1 - m_k^2) w_{ik}^2 - 4 \sum_{k \in e} r(e_k) r(-e_k) w_{ki}^2 \right)
\]

\[
-4 \sum_{k \in e, j \in s} r(e_k) r(-e_k) m_{ij} w_{kj} w_{ki}
\]
Interpretation of different interaction terms appearing in Eq. 1.26. The open and shaded nodes are hidden and evidence nodes, respectively. Solid arrows indicate the graphical structure in the network. Dashed arrows indicate interaction terms that appear in Eq. 1.26.

\[
+2 \sum_{k \in s, j \in r} (1 - m_i^2) r(-c_k) e_j w_{ij} w_{ki}
\]  

with \(m_i = m_i^d \approx m_i^p, i \in s\) and \(r\) given by Eq. 1.24.

The different terms that appear in this equation can be easily interpreted. The first term describes the lowest order forward influence on node \(i\) from its parents. Parents can be either evidence or hidden nodes (fig. 1.3a). The third term describes to lowest order the effect of Bayes' rule: it affects \(m_i\) such that the observed evidence on its children becomes most probable (fig. 1.3b). Note, that this term is absent when the evidence is as expected under model \(r: r(e_k) = 1\). The fourth and fifth terms are the quadratic contributions to the first and third terms, respectively. The sixth term describes 'explaining away'. It describes the effect of hidden node \(j\) on node \(i\), when both have a common observed child \(k\) (fig. 1.3c). The last term describes the effect on node \(i\) when its grandchild is observed (fig. 1.3d).

Note, that these equations are different from Eq. 1.17. When one applies Eq. 1.17 to sigmoid belief networks, one requires additional approximations to compute \(\log \sigma(x, h_i)\) [2].

Since only feed-forward connections are present, one can order the nodes such that \(w_{ij} = 0\) for \(i < j\). Then the first order mean field equations can be solved in one single sweep starting with node 1. The full second order equations can be solved by iteration, starting with the first order solution.
Figure 1.4
Comparison Saul's variational method and first and second order mean field methods on Saul's 2 x 4 x 6 toy problem. Figures show histograms of RMS errors of mean node activities over 1000 runs. Top left: Saul's variational method. Top right: first order mean field approximation. Bottom left: second order mean field (TAP) approximation. Bottom right: difference RMS(Saul)-RMS(TAP). Mean RMS errors are 0.0161 (Saul), 0.048 (MF) and 0.010 (TAP), respectively.

We illustrate Eq. 1.26 on Saul's (2 x 4 x 6) toy problem defined [2]. The top layer with 2 nodes is fully connected to the second layer with 4 nodes, which is fully connected to the bottom layer with 6 nodes. In the toy problem, the weights (w') and thresholds (θ') are drawn randomly from [-1, 1]. Because our definition of σ(x) differs from Saul's (σ\text{Saul}(x) = σ(x/2)) and because we use ±1 coding instead of 0, 1 coding, we must convert: w_{ij} = w'_{ij}/4 and θ_i = θ'_i/2 + \sum_j w'_{ij}/4.

We randomly generate 1000 networks and compute the mean rates using Saul's method (i.e. Eq. 1.17 plus an additional approximation to compute the expectation value of the logarithm), the first order mean field approximation (MF), the second order mean field approximation (TAP) and the exact method. We compute the RMS errors in the mean rates of the three approximate methods (√{\frac{1}{n} \sum_i (m^x_i - m^\text{approx}_i)^2}) for each run. In fig. 1.4 we plot the histograms for both Saul's method and the mean field methods. In addition, we plot the histogram of the difference in error between Saul's method and the TAP result. We conclude that the TAP approximation gives significantly better estimates of the mean rates of the hidden variables than Saul's method and requires less time (0.006 versus 10.5 seconds on Linux 300 MHz PC).
Discussion

In this chapter, we have proposed a novel method for the computation of mean field approximations. The approach is more general than the usual mean field approach, since it does not require the notion of a partition function. It is based on a Taylor series expansion of the marginal probabilities, rather than on the extremisation of a bound.

We have applied this approach to specific classes of probability distributions, such as Boltzmann machines, asymmetric stochastic neural networks, directed graphical models and sigmoid belief networks. We have shown, that this approach is applicable in all these cases, and gives reliable approximations.

An important issue is when to expect the Taylor series approximation to give reliable results. In the case of Boltzmann distributions, this issue was addressed in [14]. It can be shown, that the radius of convergence of the Taylor series can be computed approximately within the mean field framework. This argument however, assumes that the mean field approximation is reliable, which we know is not true for all probability distributions. Therefore, whether this approach can be extended to arbitrary probability distributions remains an open question.

It has been reported [15, 16] that improved approximations can also be obtained by using a very different approach, which is to use an approximating distribution \( q \) that is not factorized, but still tractable. A promising way to proceed is therefore to combine both approaches and to do a second order expansion around a manifold of distributions with non-factorized yet tractable distributions. In this approach the sufficient statistics of the tractable structure is expanded, rather than the marginal probabilities of single variables.

A The TAP equation for Boltzmann distributions and asymmetric networks

Define \( h_i = \sum_{j=0}^n w_{ij} x_j, x_0 = 1 \) and \( w_{i0} = \theta_i, t_i = \tanh h_i \). First, we compute the derivatives in the factorized point \( q \) using Eq. 1.10. This means that after differentiation, we set \( w_{ij} = 0, p(x) \to q(x) \) and \( \tanh h_i \to m_i \). Subscripts \( ij \) after a comma denote differentiation with
Mean field theory for graphical models

respect to \( w_{ij} \).

\[
m_{i,jk} = \frac{\partial m_i}{\partial w_{jk}} = \left( \sum_x p(x) \tanh h_i + p(x)(1 - \tanh^2 h_i)x_k \delta_{ij} \right) |_q
\]

\[= (1 - m_i^2) \delta_{ij} m_k \quad (1.27)
\]

\[
m_{i,jk',r'} = (1 - m_i^2) \left( m_{k,jk'} \delta_{ij} + m_{k',jk} \delta_{i,j'} - 2m_i \langle x_k x_{k'} \rangle \delta_{ij} \delta_{i,j'} \right)
\]

\[= (1 - m_i^2) \left( (1 - m_k^2) \delta_{ij} \delta_{i,r'} + (1 - m_i^2) \delta_{jk} \delta_{i,j'} \right)
\]

\[- 2m_i \langle x_k x_{k'} \rangle \delta_{ij} \delta_{i,j'} \quad (1.28)
\]

Substituting these derivatives in Eq. 1.9, we obtain

\[
0 = \Delta m_i = \sum_k m_k \Delta w_{ik} + \sum_{kk'} (1 - m_k^2) m_{k'} \Delta w_{ik} \Delta w_{kk'}
\]

\[= m_i \sum_{kk'} \langle x_k x_{k'} \rangle \Delta w_{ik} \Delta w_{kk'} \quad (1.29)
\]

The first order term gives \( \sum_k m_k \Delta w_{ik} = 0 + \mathcal{O}(\Delta^2) \). Therefore, the second term is of \( \mathcal{O}(\Delta^3) \) and can be ignored. In the third term we can make a similar argument, so that the final result is:

\[
0 = \Delta m_i = \sum_k m_k \Delta w_{ik} - m_i \sum_{k>0} \Delta w_{ik}^2 (1 - m_k^2)
\]

\[= \theta_i^0 - \theta_i^0 + \sum_{k>0} w_{ik} m_k - m_i \sum_{k>0} w_{ik}^2 (1 - m_k^2) \quad (1.30)
\]

By substituting \( \theta_i^0 = \tanh^{-1} m_i \) we obtain Eq. 1.11.

References

SNN-97-001, F-97-005.