Probabilistic data modeling with adaptive TAP mean field theory

Manfred Opper
Neural Computing Research Group,
Department of Computer Science and Applied Mathematics,
Aston University, Birmingham B4 7ET, United Kingdom

Ole Winther
Center for Biological Sequence Analysis, BioCentrum,
Technical University of Denmark, B208, 2800 Lyngby, Denmark.

July 17, 2001

Abstract

We demonstrate for the case of single layer neural networks how an extension of the TAP mean field approach of disorder physics can be applied to the computation of approximate averages in probabilistic models for real data.

1 Introduction

Presently there is a growing interest within the *machine learning* and *artificial intelligence* communities to apply approximation techniques from statistical physics to probabilistic data models. Such models explain complex observed data by a set of unobserved, *hidden* random variables based on their joint distribution. Examples are: Bayes belief networks (used as trainable expert systems), independent component analysis (abbreviated ICA, which detects independent sources in nonlinear signal processing), Gaussian process models (modelling hidden spatial structures by random fields) and Boltzmann machines (the Ising version of the random fields). Learning and inference about the hidden variables requires the computation of averages that are quite similar to the ones that are encountered in equilibrium statistical physics. For large models of practical interest, the vast increase of computational complexity precludes exact calculations. Already simple mean field methods which neglect all correlations between random variables were found to give often qualitatively good results in such cases.

Since dependencies between the variables are often long ranged and fluctuating, one may expect that the *TAP* mean field method[1] developed for disordered systems is a good method to improve over the simple MF method for probabilistic data models. For applications of this idea see [3, 4, 5, 6, 7, 8, 9] and the review in [10].

Unfortunately, the Onsager correction to the simple MF theory provided by the TAP approach–assuming models with extensive connectivity–will explicitly depend on the probability distribution of the couplings between random variables. While in the models of statistical physics such distributions are given explicitly, they are usually unknown in the data models. To overcome this problem, we have developed a version of the TAP method [11, 12] which adapts the Onsager term to the concrete set of couplings. We will demonstrate the method for the case of learning with a single layer neural network and show its significance compared to an approach which simply assumes a given distribution of data.

Our approach was developed for models with pairwise interactions between variables S_i, $i = 1, \ldots, N$

$$P(\mathbf{S}) = \frac{\rho(\mathbf{S})}{Z(\boldsymbol{\theta}, \mathbf{J})} \exp\left[\sum_{i<j} S_i J_{ij} S_j + \sum_i S_i \theta_i\right] . \tag{1}$$

Here $\mathbf{S} = (S_1, \ldots, S_n)$ denotes the set of hidden variables. All self-interactions are contained in the factorizing distribution $\rho(\mathbf{S}) \equiv \prod_j \rho_j(S_j)$ which also contains all single variable constraints of the variables S_i like their *range*, their *discreteness* etc. Examples of models that are included in this framework are Ising models (like the SK model, the Hopfield model, the Boltzmann machine in the neural computation context), the finite temperature versions of the matching and travelling salesman problems [13, 14], Gaussian process models [9] and the ICA model of [15]. We will show next, how a single layer network fits in to this framework.

2 The Perceptron

Perceptrons are single layer neural networks which are parametrized by a vector of weights $\mathbf{w}$. Perceptrons model the output y to an input vector $\mathbf{x} \in R^N$ as $y = \mathbf{w} \cdot \mathbf{x}$ for regression and $y = \text{sign } \mathbf{w} \cdot \mathbf{x} = \pm 1$ for binary classification. To estimate reasonable values of the unknown (hidden) variable $\mathbf{w}$ from an observed training set $D = \{(\mathbf{x}_k, y_k), k = 1, \ldots, m\}$ within a probabilistic model, one defines a probability $P(y|\mathbf{w} \cdot \mathbf{x})$ for the observations y given inputs $\mathbf{x}$ and weights $\mathbf{w}$. For classification we use $P(y|\mathbf{w} \cdot \mathbf{x}) = \phi(y \frac{\mathbf{w} \cdot \mathbf{x}}{\sigma})$, where $\phi(z) \equiv \int_{-\infty}^{z} Dt$ and $Dt = e^{-t^2/2} dt/\sqrt{2\pi}$. In the noise-free limit ϕ reduces to the unit step-function. For regression with additive Gaussian noise the likelihood is $P(y|\mathbf{w} \cdot \mathbf{x}) \propto e^{-(y - \mathbf{w} \cdot \mathbf{x})^2/2\sigma^2}$. If we assume

[1] TAP was first introduced by Thouless, Anderson and Palmer (TAP) [1] to treat the Sherrington-Kirkpatrick (SK) model of disordered magnetic materials [2].

a *prior* distribution over the couplings $P(\mathbf{w}) = \prod_{i=1}^{N} \rho_i(w_i)$, we can use Bayes theorem to compute the posterior distribution

$$P(\mathbf{w}|D) = \frac{1}{Z} \prod_{i=1}^{N} \rho_i(w_i) \prod_{j=1}^{m} P(y_j|\mathbf{w} \cdot \mathbf{x}_j) \ . \tag{2}$$

We may now take the averaged weights $\langle \mathbf{w} \rangle$ over the distribution eq. (2) in the predictions for new data, i.e. predict on $\mathbf{x}$: $\langle \mathbf{w} \rangle \cdot \mathbf{x}$ for regression and sign $\langle \mathbf{w} \rangle \cdot \mathbf{x}$ for classification.

To rewrite the model eq. (2) in the form eq. (1), we use the 'field theoretic' trick of introducing the fields $\sum_{i=1}^{N} x_{ki} w_i$, $k = 1, \ldots, m$ as new variables by using δ-functions and their exponential representations. Denoting the purely imaginary conjugate variables by $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_m)$, the space of variables is augmented to the set $\mathbf{S} = (\mathbf{w}, \hat{\mathbf{w}})$ where the 'prior distribution' for the hatted variables is given by

$$\hat{\rho}(\hat{w}) = \int \frac{d\hat{h}}{2\pi i} e^{-\hat{w}\hat{h}} P(y|\hat{h}) \tag{3}$$

and the augmented coupling matrix is

$$\mathbf{J}_{\mathbf{aug}} = \begin{pmatrix} 0 & \mathbf{X}^T \\ \mathbf{X} & 0 \end{pmatrix} \ ,$$

where $\mathbf{X}$ is the matrix with components x_{ik}.

3 The TAP equations

We briefly present the derivation [11, 12] of our TAP equations for the model eq. (1) based on the cavity method, assuming that we have a non-glassy system and all averages are with respect to a single state.

Defining the field $h_i = \sum_j J_{ij} S_j$, the marginal distribution of S_i can be written as

$$P_i(S) = \int \prod_{j \neq i} dS_j P(\mathbf{S}) = \frac{\rho_i(S)}{Z_i} e^{-H_i(S)} \ , \tag{4}$$

where the effective single variable Hamiltonian $H_i(S)$ is computed by averaging over all variables in a system, where S_i is left out. Denoting this 'cavity' average by $\langle \ldots \rangle_{\setminus i}$, we have $-H_i(S) = \ln \left\langle e^{Sh_i} \right\rangle_{\setminus i} = \sum_k \frac{\kappa_k^{(i)}}{k!} S^k$ where $\kappa_k^{(i)}$ are the cumulants of this *cavity* distribution. The basic assumption that all variables S_j have only weak mutual dependencies as expressed within the so–called clustering hypothesis [14] results in neglecting all cumulants $\kappa_k^{(i)}$ with $k > 2$ for fully connected systems.

Setting $V_i = \kappa_2^{(i)}$, we get

$$\langle h_i \rangle = \frac{1}{Z_i} \int dS \rho_i(S) \frac{\partial}{\partial S} e^{-H_i(S)} = \langle h_i \rangle_{\setminus i} + V_i \langle S_i \rangle \tag{5}$$

$$P_i(S) = \frac{\rho_i(S)}{Z_i} \exp\left[(\sum_j J_{ij} \langle S_j \rangle - V_i) S + \frac{1}{2} V_i S^2 \right] \tag{6}$$

for $i = 1, \ldots, N$.

The V_i's account for 'Onsager' corrections to the 'naive' MF theory. A self-consistent approximation to the V_i''s is obtained from the matrix of susceptibilities $\chi_{ij} \equiv \frac{\partial \langle S_i \rangle}{\partial \theta_j}$, which can be computed from eq. (6)[2] as $\boldsymbol{\chi} = (\boldsymbol{\Lambda} - \mathbf{J})^{-1}$, where $\boldsymbol{\Lambda} = \mathrm{diag}\{V_i + 1/\chi_{ii}\}$ is a diagonal matrix. By the *Fluctuation Dissipation Theorem*, the susceptibilities are equal to the covariances of the S_i's. Hence,

$$\langle S_i^2 \rangle - \langle S_i \rangle^2 = \frac{\partial^2 \ln Z_i}{\partial \theta_i^2} = \left[(\boldsymbol{\Lambda} - \mathbf{J})^{-1} \right]_{ii} \tag{7}$$

for $i = 1, \ldots, N$. The sets of equations (6) together with (7) are a closed sets of equations for the first and second moments of S_i which in turn enables us to approximate the full marginal distribution of S_i and the correlation functions.

4 Application to Perceptrons

We apply the TAP approach to approximate the expected weights $\langle \mathbf{w} \rangle$ for the posterior distribution (2). We consider both Ising weights $\rho(w) = \frac{1}{2}\delta(w-1) + \frac{1}{2}\delta(w+1)$ and weights with a Gaussian prior distribution $\rho(w) = e^{-w^2/2}/\sqrt{2\pi}$. In the first case, we recover the Ising result

$$\langle w_i \rangle = \tanh\left(\sum_k x_{ki} \langle \hat{w}_k \rangle - V_i \langle w_i \rangle + \theta_i \right) \tag{8}$$

and in the Gaussian case we simply get

$$\langle w_i \rangle = \sum_k x_{ki} \langle \hat{w}_k \rangle + \frac{\theta_i}{1 - V_i} \ .$$

The TAP eqs. for the hatted variables $\langle \hat{w}_k \rangle = \frac{\partial \ln \hat{Z}_0^{(k)}}{\partial \hat{\theta}_k}$ are obtained from the partition function

$$\hat{Z}_0^{(k)} = \int Dz P(y_k | \langle \hat{h}_k \rangle_{\setminus k} + \hat{\theta}_k + \sqrt{\hat{V}_k} z)$$

[2] We make the approximation that upon differentiation, the V_i's are held constant.

with $\langle \hat{h}_k \rangle_{\setminus k} = \sum_i x_{ki} \langle w_i \rangle - \hat{V}_k \langle \hat{w}_k \rangle$.

We test the significance of the adaptive TAP equations in two learning scenarios. First we check the internal consistency of the theory by comparing the cavity field calculated from the solution of the TAP equations $\langle \hat{h}_k \rangle_{\setminus k} = \sum_i x_{ki} \langle w_i \rangle - \hat{V}_k \langle \hat{w}_k \rangle$ with the 'exact' cavity field $\langle \hat{h}_k \rangle_{\setminus k}^{\rm exact}$ computed by actually *removing* example k from the training set and solving the TAP equations for the remaining $m-1$ examples and repeating this procedure for $k = 1, \ldots, m$. A precise estimate of the cavity field is of practical relevance in machine learning since it provides 'leave-one-out' estimators of the generalization error [5, 9, 11, 12]. For classification this is just the fraction of negative terms $y_k \langle \hat{h}_k \rangle_{\setminus k}$ over the training set: $\epsilon_{\rm loo} = \frac{1}{m} \sum_k \Theta(-y_k \langle \hat{h}_k \rangle_{\setminus k})$ since $\mathrm{sign} \langle \hat{h}_k \rangle_{\setminus k}$ is the leave-one-out prediction of y_k. We also compare with the usual TAP approach [3] in which the Onsager correction is computed by assuming that all components of the inputs are drawn independently at random with zero mean and unit variance.

We consider the data set, 'Sonar – Mines versus Rocks' [16] of size $m = 104$ with binary class labels $y_k = \pm 1$ and a $N = 60$ dimensional input space. We use the Gaussian prior for the weights and $\sigma^2 = 0.5$ in the likelihood. In figure 1 we plot $y_k \langle \hat{h}_k \rangle_{\setminus k}$ versus $y_k \langle \hat{h}_k \rangle_{\setminus k}^{\rm exact}$. For the adaptive theory, we find a perfect agreement between the two computations of the leave-one-out estimate: $\epsilon_{\rm loo} = \epsilon_{\rm loo}^{\rm exact} = 33/104$. For comparison, the non-adaptive TAP approach overestimates the leave-one-out error: $\epsilon_{\rm loo} = 41/104$ and $\epsilon_{\rm loo}^{\rm exact} = 33/104$.

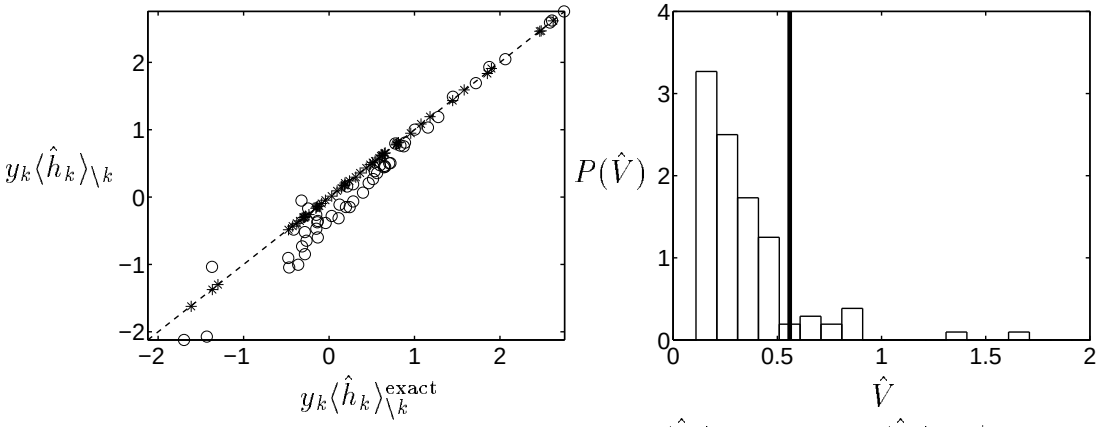

Figure 1: Test of self–consistency of TAP – $y_k \langle \hat{h}_j \rangle_{\setminus k}$ versus $y_k \langle \hat{h}_j \rangle_{\setminus k}^{\rm exact}$ for the Sonar data set. The stars/circles are for adaptive/conventional TAP. The right plot shows the distribution of the cavity variances $\hat{V}_k$. The line in the middle is the value found from the self-averaging theory.

In the second set of simulations we investigate the importance of using the correct Onsager term in the mean field equations. We consider a nontrivial regression problem with a perceptron having binary weights. See Ref. [17] for a discussion of this model in the context of communications systems. Here, we

compute an approximation to the free energy $\Phi = -\ln Z$ based on the TAP approach (details of the calculation are given in [12]). In probabilistic data models, the free energy usually equals the negative log Likelihood of the observed data, i.e. $\Phi = -\ln P(\mathbf{y})$, which can be used for deciding which model parameters (e.g. noise rate) gives the best fit to data.

Since the regression model is quadratic in the variables w_i, we can also obtain an alternative representation in the form of eq. (1) without introducing auxiliary variables. We have $\prod_k P(y_k|\mathbf{w}\cdot\mathbf{x}_k) \propto e^{\sum_{i>j} w_i J_{ij} w_j + \sum_i w_i \theta_i}$. The couplings and external fields are given by $J_{ij} = -\sum_k x_{ki}x_{kj}/\sigma^2$ and $\theta_i = \sum_k x_{ki}y_k/\sigma^2$. For this Ising model, we test the performance of TAP equations with Onsager terms that are based on the completely wrong assumptions that the random matrices $\mathbf{J}$ correspond to the SK- and Hopfield-models respectively [14, 12].

In the simulations we set $N = 60$, $\sigma^2 = 0.2$ and generate a training set using a noise-free binary 'teacher' perceptron: $y = \mathbf{T}\cdot\mathbf{x}$ with $T_i = \pm 1$. The x_{ki}'s are i.i.d with zero means and unit variance. Since the distribution of couplings is known, we also give the results of the conventional TAP approach, where the Onsager term is calculated for that distribution [3, 5].

Figure 2 shows the TAP mean field free energy Φ found in simulations using different expressions for Onsager terms together with the prediction of an analytical replica calculation [18]. The simulations are averaged over 100 runs and the error-bars are of the size of the symbols. Both the adaptive method and the one that uses the Onsager term based on the correct input distribution are in excellent agreement with the replica result. Using the SK and Hopfield Onsager terms lead to completely wrong estimates of the free energy.

5 Summary and Outlook

For the example of single layer neural networks we have shown that a generalization of the TAP approach for disordered systems can be successfully applied to adaptive data models. We have demonstrated that TAP equations which are based on false assumptions about disorder distributions can lead to unreliable approximations for *leave-one-out estimators* of the model's prediction errors and for *free energies* which can serve as a practical tools for validating the models and their predictions. On the other hand, our *adaptive* TAP approach avoids such assumptions by computing the Onsager correction self-consistently from the data.

At present we are working on the development of efficient algorithms for solving TAP equations which allow for an application of our TAP method to a variety of practically relevant data models. Of special importance are further *sparse approximations* which reduce the complexity of matrix operations when the number of variables is very large. Such approximations are based on optimally projecting probability distributions onto subspaces that are spanned by a smaller

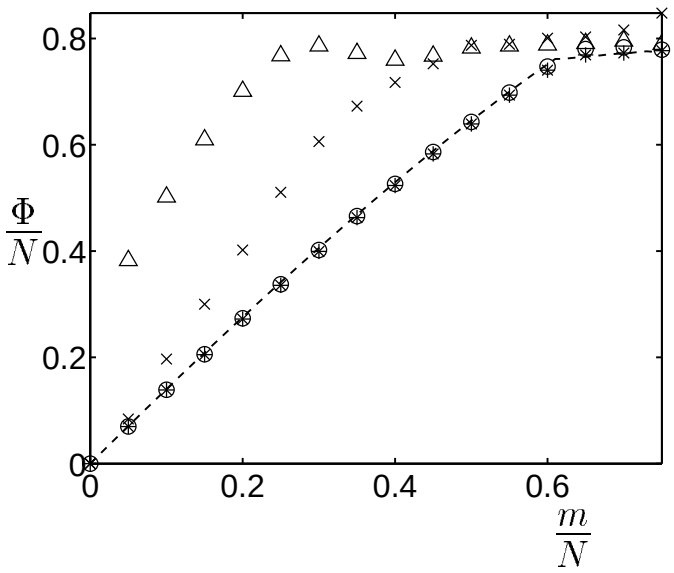

Figure 2: The free energy Φ as a function of the training set size m. The dashed line is the prediction of replica theory. Stars/circles (almost coinciding) are the results for adaptive TAP/correct distribution TAP. Crosses/triangles are the results for TAP with the Hopfield/SK Onsager term.

set of representative variables.

Further research is necessary to understand if the limitations of our method to models with 'non-glassy' behaviour can be overcome.

Acknowledgments

This research is supported by the Danish Research Councils through the Center for Biological Sequence Analysis.

References

[1] D. J. Thouless, P. W. Anderson and R. G. Palmer, Phil. Mag. **35**, 593 (1977).

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

[3] M. Mézard, J. Phys. A (Math. Gen.) **22**, 2181 (1989).

[4] K. Y. M. Wong, Europhys. Lett. **30**, 245 (1995).

[5] M. Opper and O. Winther, Phys. Rev. Lett. **76**, 1964 (1996).

[6] H. J. Kappen and F. B. Rodríguez, Neural Computation **10**, 1137 (1998).

[7] T. Tanaka, Phys. Rev. E **58**, 2302 (1998).

[8] Y. Kabashima and D. Saad, Euro. Phys. Lett. **44**, 668 (1998).

[9] M. Opper and O. Winther, Neural Computation **12**, 2655 (2000).

[10] M. Opper and D. Saad, eds. *Advanced Mean Field Methods, Theory and Practice*, MIT Press (2001).

[11] M. Opper and O. Winther, Phys. Rev. Lett. **86**, 3695 (2001).

[12] M. Opper and O. Winther, Adaptive and Self-averaging Thouless-Anderson-Palmer Mean Field Theory for Probabilistic Modeling, Phys. Rev. E *accepted* (2001).

[13] M. Mézard and G. Parisi, J. Phys. Lett. **46**, L771 (1995).

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

[15] D. J. C. MacKay: *Maximum Likelihood and Covariant Algorithms for Independent Component Analysis*, University of Cambridge, Cavendish Laboratory, Draft 3.7, (1996).

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

[17] T. Tanaka, to appear in (NIPS'2000), MIT Press (2001).

[18] H. S. Seung, H. Sompolinsky and N. Tishby, Phys. Rev. A. **45**, 6056 (1992).