Validity of TAP equations in Neural Networks

M.A.R. Leisink* and H.J. Kappen†
Department of Biophysics
University of Nijmegen, Geert Grooteplein 21
NL 6525 EZ Nijmegen, The Netherlands ‡

May 18, 1999

Abstract

The statistics of a Boltzmann machine can be approximated using the TAP equations combined with linear response theory. We discuss the validity of the TAP equations, in particular for finite size networks. We present an algorithm that determines if a particular solution of the TAP equations is valid.

1 Introduction

Boltzmann machines are networks of stochastic binary variables (neurons). All neurons s_i are linked to each other with symmetric weights $w_{ij} = w_{ji}$. Due to this symmetry the probability distribution is given by the Boltzmann-Gibbs distribution which is a known function of the weights and thresholds of the network [1].

Since the exact computation of the statistics is intractable, one has to make an approximation. Plefka [2] presented an elegant way to derive an approximation (originally found by Thouless, Anderson and Palmer [3]) called the TAP equations. The method is based on a small weight expansion around a tractable, decoupled network and is an extension of the naive mean field method.

The small weight expansion only converges within the radius of convergence. Outside that radius expansions upto any order give a poor approximation. Therefore, the TAP expansion is only valid if the weights and the TAP solution are within that radius. Plefka derived some conditions for the convergence, but they can only be used in the limit of an infinite size network. In section 3 we derive the conditions for a finite size network, which is a more realistic case for neural networks.

In section 4 we illustrate the validity condition numerically by computing correlations $\langle s_i s_j \rangle$ both exactly and using the TAP scheme. In addition we show the results of Boltzmann machine learning, where the TAP approximation is used for the needed statistics [4][5]. The validity of the TAP solution is computed after each weight update.

*http://www.mbfys.kun.nl/~martijn
†http://www.mbfys.kun.nl/~bert
‡This research is supported by the Technology Foundation STW, applied science devision of NWO and the technology programme of the Ministry of Economic Affairs.

2 Theory

We consider a network of N neurons, $s_i = \pm 1$, with thresholds θ_i and symmetric weights $w_{ij} = w_{ji}$. The energy of such a network is given by

$$E(\vec{s}, \alpha) = -\sum_i \theta_i s_i + \alpha E_{\text{int}} \tag{1}$$

where E_{int} stands for the interaction energy defined by

$$E_{\text{int}}(\vec{s}) = -\frac{1}{2}\sum_{ij} w_{ij} s_i s_j \tag{2}$$

The probability to find the system in a state $\vec{s}$ is given by

$$P(\vec{s}, \alpha) = \exp\left(-E(\vec{s}, \alpha) - \Psi(\alpha)\right) \tag{3}$$

where $\Psi(\alpha)$ is a normalisation constant defined by

$$\Psi(\alpha) = \log \sum_{\text{all } \vec{s}} \exp\left(-E(\vec{s}, \alpha)\right) \tag{4}$$

which is minus the well known free energy.

This free energy is a function of the independent variables θ_i and w_{ij}. We perform a Legendre transformation to make

$$m_i \stackrel{\text{def}}{=} \frac{\partial \Psi}{\partial \theta_i} \tag{5}$$

the new independent variables instead of θ_i. Hence, we obtain the Legendre transform of Ψ

$$\Phi(m_i, w_{ij}, \alpha) = \sum_i \theta_i m_i - \Psi(\theta_i, w_{ij}, \alpha) \tag{6}$$

where m_i and w_{ij} are the independent variables and θ_i is a function of them defined by

$$\theta_i = \frac{\partial \Phi}{\partial m_i} \tag{7}$$

We expand $\Phi(m_i, w_{ij}, \alpha)$ in α

$$\Phi(\alpha) = \Phi(0) + \alpha \Phi'(0) + \frac{1}{2}\alpha^2 \Phi''(0) + \mathcal{O}(\alpha^3) \tag{8}$$

where a prime denotes differentiation with respect to α. We directly obtain from [2]

$$\Phi'(\alpha) = \langle E_{\text{int}} \rangle_\alpha \tag{9}$$

$$\Phi''(\alpha) = \langle E_{\text{int}} \rangle_\alpha^2 - \langle E_{\text{int}}^2 \rangle_\alpha + \left\langle E_{\text{int}} \sum_i \frac{\partial \theta_i}{\partial \alpha}(s_i - m_i) \right\rangle_\alpha \tag{10}$$

Evaluating these expressions at $\alpha = 0$ gives

$$\Phi(0) = \sum_i \left\{ \frac{1+m_i}{2} \log \frac{1+m_i}{2} + \frac{1-m_i}{2} \log \frac{1-m_i}{2} \right\} \tag{11}$$

$$\Phi'(0) = -\frac{1}{2}\sum_{ij} w_{ij} m_i m_j \tag{12}$$

$$\Phi''(0) = -\frac{1}{4}\sum_{ij} w_{ij}^2 \left(1 - m_i^2\right)\left(1 - m_j^2\right) \tag{13}$$

We find the TAP approximation for Φ by substituting equations 11 to 13 in equation 8 and setting α to one.

To find the value for m_i, we use the property of the Legendre transformation as in equation 7

$$\theta_i = \frac{\partial \Phi}{\partial m_i} = \tanh^{-1} m_i - \alpha \sum_j w_{ij} m_j + \alpha^2 m_i \sum_j w_{ij}^2 \left(1 - m_j^2\right) \tag{14}$$

which we recognise as the TAP equations for $\alpha = 1$. The correlations are given by (see also [4])

$$\langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle = \frac{\partial \Psi^2}{\partial \theta_i \partial \theta_j} \stackrel{\text{def}}{=} \chi_{ij} \tag{15}$$

where the inverse of the matrix χ is given by

$$\left(\chi^{-1}\right)_{ij} = \left(\frac{1}{1 - m_i^2} + \alpha^2 \sum_k w_{ik}^2 \left(1 - m_k^2\right)\right)\delta_{ij} - \alpha w_{ij} - 2\alpha^2 w_{ij}^2 m_i m_j \tag{16}$$

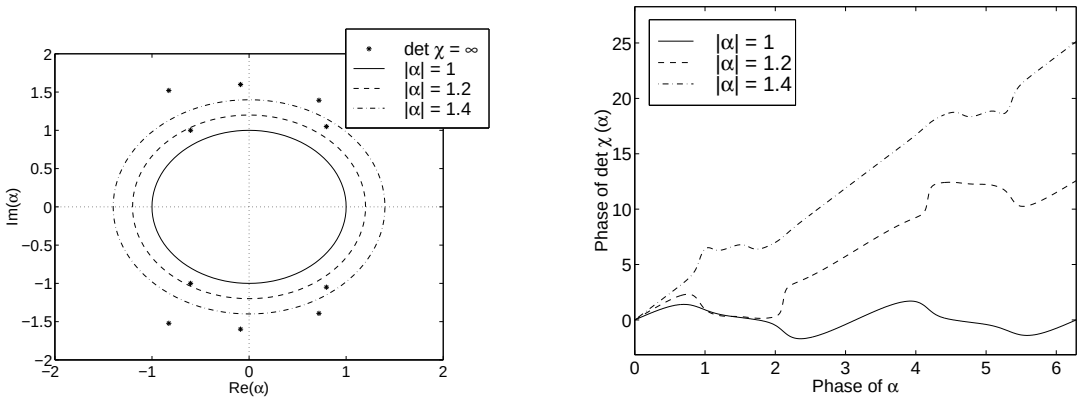

Figure 1: The left figure shows the complex α-plane and the singularities of $\det \chi(\alpha)$ for a network of five neurons. The weights were randomly chosen from a Gaussian with $\sigma = 0.45$ and $\mu = 0$. The mean field variables m_i are zero, which is a solution of the TAP equations since the thresholds are chosen zero. The right figure shows the increase in phase of $\det \chi(\alpha)$ for three different radii $|\alpha|$. Since the circles $|\alpha| = 1.2$ and $|\alpha| = 1.4$ enclose poles, there is a net increase of the phase of $\det \chi(\alpha)$ and the solution $m_i = 0$ is invalid.

since

$$\frac{\partial^2 \Psi}{\partial\theta_i \partial\theta_j} = \frac{\partial m_i}{\partial\theta_j} = \left(\frac{\partial\theta}{\partial m}\right)^{-1}_{ij} = \left(\frac{\partial^2 \Phi}{\partial m^2}\right)^{-1}_{ij} \quad (17)$$

3 Validity of the TAP expansion

Let the radius of convergence of the expansion in equation 8 be ρ. For $\alpha < \rho$ the error of the TAP approximation is $\mathcal{O}(\alpha^3)$. However, if $\alpha > \rho$ the expansion does not converge and any truncation of the Taylor series is meaningless. Moreover, the addition of an extra expansion term will in general increase the error instead of giving a better approximation. Since we set $\alpha = 1$ to obtain the TAP approximation, we require $\rho > 1$ [2].

For an exact Boltzmann machine, we derive

$$\frac{\partial \Phi}{\partial \alpha} = \frac{1}{2}\sum_{ij} w_{ij} m_i m_j + \frac{1}{2}\sum_{ij} w_{ij}\chi_{ij} \quad (18)$$

We use the fact that ρ is the same for $\Phi(\alpha)$ and $\partial\Phi/\partial\alpha$. Furthermore ρ is equal to the distance between the origin and the nearest singular point in the complex α-plane. Thus the singularities of the matrix $\chi(\alpha)$ given the thresholds, weights and mean field variables m_i determine the radius of convergence ρ.

To find these singularities we assume that the approximation for $\chi(\alpha)^{-1}$, given by equation 16 is good within the radius of convergence and hence may be used to find the singularities. Plefka [2] showed that this assumption is correct for the SK-model [6] in the limit of infinite networks.

One should keep in mind that a direct computation of the approximated $\partial\Phi/\partial\alpha$ to obtain $\chi(\alpha)$ will never give any poles, since Φ is a Taylor expansion (i.e. a polynomial function of α). A solution is to use $\chi^{-1}(\alpha)$ as in (16), which may be not of maximum rank, so that $\chi(\alpha)$ does have poles. This is certainly not a unique choice, but it appears to be quite good according to our simulations.

Consider the circle $C : |\alpha| = 1$ in the complex α-plane. This circle is mapped to a closed curve by the map $\det \chi(\alpha)$ Since this is a analytic function except for a finite number of poles, the integral

$$\frac{1}{2\pi \mathrm{i}} \oint_C \det \chi(\alpha) \, \mathrm{d}\alpha \tag{19}$$

is equal to the number of poles within C. Thus the increase in phase of $\det \chi(\alpha)$, when α follows C, gives the number if poles bounded by $|\alpha| = 1$. This is shown in figure 1. Thus the validity condition $\rho > 1$ corresponds to a zero integral in equation 19.

The calculation of the determinant is $\mathcal{O}(N^3)$. The increase of phase of $\det \chi(\alpha)$ is somewhere between zero and $2\pi N$. Therefore, in the worst case of a maximum increase, the step size with which we increment the phase of α must be $\mathcal{O}(N^{-1})$ to be able to compute this phase change with enough accuracy. Hence the computational complexity of the algorithm is somewhere between $\mathcal{O}(N^3)$ and $\mathcal{O}(N^4)$.

4 Results

We initialise a network of $N = 14$ neurons with weights drawn from a Gaussian with standard deviation $1/\sqrt{N}$ and zero mean (which is the so called SK-model [6]). The network has its thresholds set to zero and therefore $m_i = 0$ is always a solution of equation 14. It is important to understand that although the solution $m_i = 0$ is stable and corresponds to the exact $\langle s_i \rangle = 0$, the TAP expansion is meaningless when it does not converge. As a consequence one can expect large errors in, for instance, the approximated correlations. Therefore we still need to know the validity of the solution.

We multiply all weights with a scaling factor which we vary from zero to two. For each value of the scaling factor we compute the correlations using equation 15 with the solution $m_i = 0$. In figure 2 we have plotted the approximation error of the correlations defined by

$$\eta = \frac{1}{2} \sum_{ij} \left(\langle s_i s_j \rangle_{\text{exact}} - \langle s_i s_j \rangle_{\text{TAP}} \right)^2 \tag{20}$$

versus the scaling of the weights. One can see an enormous increase of the error starting roughly at the point that the TAP solution is invalid according to our algorithm.

To understand the use of the validity condition in Boltzmann machine learning, we train a network of eight neurons using the TAP approximation with linear response as in [4]. The target distribution is the Asia problem, where the correlations between some diseases and findings are modelled [7]. This results in a probability distribution of eight binary neurons, which we try to learn without hidden units. Learning was done using the gradient descent rule [1]

$$\Delta\theta_i = \eta \left(\langle s_i \rangle_{\text{asia}} - \langle s_i \rangle_{\text{net}} \right) \tag{21}$$

$$\Delta w_{ij} = \eta \left(\langle s_i s_j \rangle_{\text{asia}} - \langle s_i s_j \rangle_{\text{net}} \right) \tag{22}$$

with a learning rate $\eta = 0.05$. $\langle \cdot \rangle_{\text{net}}$ and $\langle \cdot \rangle_{\text{asia}}$ are the averages in the current network and those in the Asia problem, respectively.

The Kullback divergence between the target and the learned distribution is plotted at the left of figure 3 for both the exact and the TAP learning procedure. Note that the Kullback error is not available in large problems due to the computational intractability. At the right the number of poles is plotted for each learning step. As one can see the Kullback divergence generally decreases if the number of poles is zero, but increases dramatically if not. In the latter case the TAP solution is wrong, since the TAP expansion does not converge.

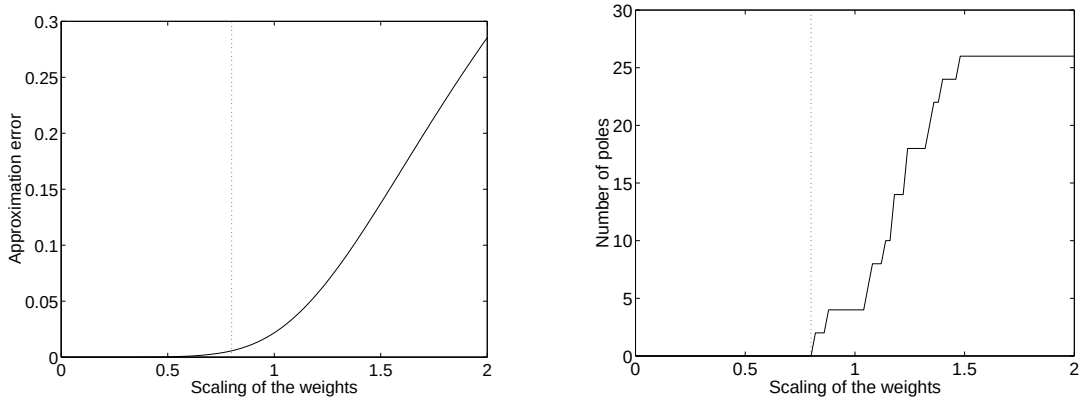

Figure 2: The left graph shows the error of the approximated correlations versus the scaling of the weights. Beyond a scaling of 0.8 the TAP solution $m_i = 0$ is not valid since the number of poles, as is shown in the right graph, is greater than zero.

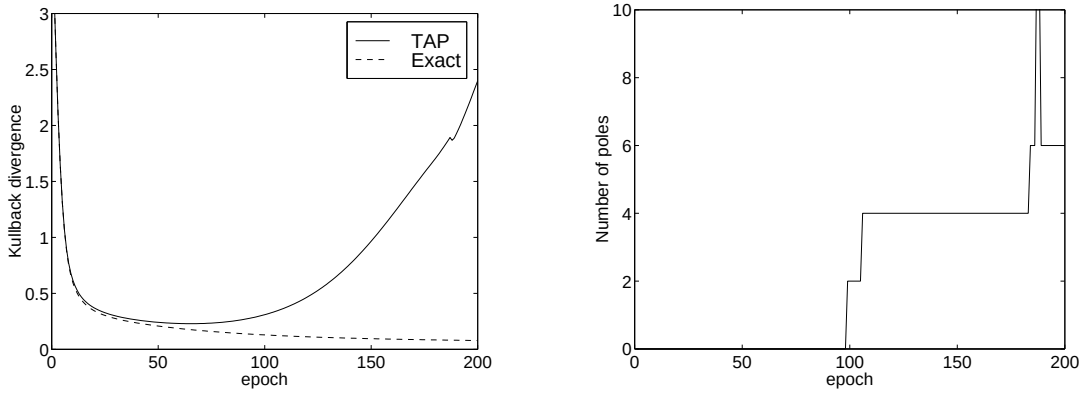

Figure 3: The learning of the Asia problem. The left graph shows the Kullback divergence for the exact and the approximated learning. The right graph shows the number of poles of $\det \chi(\alpha)$ with $|\alpha| < 1$ for each learning step. As the number of poles is greater than zero, the TAP solution is wrong and learning should be stopped at that point.

5 Conclusions

We have presented an algorithm to determine the validity of a TAP solution. The computational complexity of the algorithm is polynomial in the size of the network. We have shown that the correlations are badly approximated if the solution of the TAP equations is invalid according to the algorithm.

Furthermore we have applied the algorithm to Boltzmann machine learning. There are targets for which the TAP solution reaches the invalid region after some epochs. Therefore it is reasonable to believe that such a target lies in the invalid region. We have shown that in this region the learning procedure in general increases the Kullback divergence and thus decreases the network performance. If the invalid region is entered, one can decide either to stop learning and use the realisation of the network so far or to mark the problem as unsolvable within the TAP approximation.

References

[1] D. Ackley, G. Hinton, and T. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

[2] T. Plefka. Convergence condition of the TAP equation for the infinite-ranged ising spin glass model. *J.Phys.A: Math.Gen.*, 15:1971–1978, 1981.

[3] D.J. Thouless, P.W. Andersson, and R.G. Palmer. Solution of 'solvable model of a spin glass'. *Philisophical Magazine*, 35(3):593–601, 1977.

[4] H.J. Kappen and F.B. Rodríquez. Efficient learning in Boltzmann machines using linear response theory. *Neural Computation*, 10(5):1137–1156, 1998.

[5] H.J. Kappen and F.B. Rodríquez. Boltzmann machine learning using mean field theory and linear response correction. In M.I. Jordan, M.J. Kearns, and S.A. Solla, editors, NIPS, 1998.

[6] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Physical Review Letters*, 35(26):1793–1796, 12 1975.

[7] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Statistics Society B*, 50(2):157–194, 1988.