Learning Higher Order Boltzmann Machines using Linear Response

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

March 23th, 1998

Abstract

Boltzmann machines are able to represent some probability distribution but the exact learning algorithm needs a time that is exponential in the number of neurons. The approximation method called Linear Response is not only applicable to machines with only second order interactions, but can be extended to Boltzmann machine with third and higher order interactions. It is shown that this can be used to estimate probability distributions which have strong third or higher order correlations.

1 Introduction

A Boltzmann machine is a network of stochastic variables (neurons), which value is either plus or minus one. All neurons are linked to each other with symmetric weights $w_{ij} = w_{ji}$. Due to this symmetry the stationary probability distribution is given by the Boltzmann-Gibbs distribution $P(\vec{s})$, which is a known function of the weights and thresholds of the network [1].

Since the computation of this distribution requires an amount of time proportional to 2^N, where N is the number of neurons, an approximation is needed. Kappen and Rodríguez [2] have shown an approximation method called Linear Response which is an order N^3 algorithm.

We extend this approximation to higher order networks. These higher order networks have not only first and second order interactions θ_i and w_{ij} but also third and higher order interactions like w_{ijk} or w_{ijklm}.

We show that it is possible to derive do the Linear Response approximation similar to the second order case. Moreover in the absence of hidden units it is possible to obtain an immediate approximation of the weights instead of learning the machine using gradient descent.

---

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

2 Approximation of the Free Energy

A higher order Boltzmann machine has a probability distribution

$$P(\vec{s}) = \frac{1}{Z}\exp(-E(\vec{s})) \tag{1}$$

where

$$Z = \sum_{\text{all } \vec{s}} \exp(-E(\vec{s})) \tag{2}$$

and

$$E(\vec{s}) = -\sum_i \theta_i s_i + \alpha E_{int}(\vec{s}) \tag{3}$$

The interaction energy in the last equation is given by all interactions in the system

$$E_{int}(\vec{s}) = -\frac{1}{2}\sum_{ij} w_{ij} s_i s_j - \frac{1}{6}\sum_{ijk} w_{ijk} s_i s_j s_k - \dots \tag{4}$$

For $\alpha = 1$ we have the fully connected Boltzmann machine; for $\alpha = 0$ the model is decoupled.

Calculation of the free energy $F = -\log Z$ requires a time that is exponential in the number of neurons. Therefore we need an approximation. We will follow the work of Plefka [3] and expand this free energy in α, since the model is calculable for $\alpha = 0$.

First we introduce new variables

$$m_i = \langle s_i \rangle = -\frac{\partial F}{\partial \theta_i} \tag{5}$$

and perform a standard Legendre transformation to make these m_i the new independent variables.

$$G(\vec{m}, \vec{w}, \alpha) = F(\vec{\theta}, \vec{w}, \alpha) + \sum_i \theta_i m_i \tag{6}$$

where G is known as the Gibbs free energy which has m_i and the weights as independent variables.

We approximate this G by a Taylor expansion around $\alpha = 0$

$$G(\vec{m}, \vec{w}, \alpha) \approx G(\vec{m}, \vec{w}, 0) + \alpha \left.\frac{\partial G}{\partial \alpha}\right|_{\alpha=0} \tag{7}$$

Since we know that $G(\vec{m}, \vec{w}, 0)$ represents the Gibbs potential of the decoupled spins, we obtain

$$\begin{aligned} G(\vec{m}, \vec{w}, \alpha) \quad \approx \quad & \sum_i \left\{ \frac{1+m_i}{2}\log\frac{1+m_i}{2} + \frac{1-m_i}{2}\log\frac{1-m_i}{2} \right\} \\ & + \alpha \langle E_{int} \rangle_{\alpha=0} \end{aligned} \tag{8}$$

and the mean field variables m_i are calculated by using the property of the Legendre transformation that $\theta_i = \partial G / \partial m_i$, where we put in the approximated Gibbs free energy.

3 Estimating the Correlations

Notice that in the exact case the correlations can be expressed as higher order derivatives of the free energy

$$\begin{aligned}
\langle s_i \rangle &= -\frac{\partial F}{\partial \theta_i} \\
\langle s_i s_j \rangle &= -\frac{\partial^2 F}{\partial \theta_i \partial \theta_j} + \langle s_i \rangle \langle s_j \rangle \\
\langle s_i s_j s_k \rangle &= -\frac{\partial^3 F}{\partial \theta_i \partial \theta_j \partial \theta_k} + \langle s_i \rangle \langle s_j s_k \rangle + \langle s_j \rangle \langle s_k s_i \rangle \\
&\quad + \langle s_k \rangle \langle s_i s_j \rangle - 2 \langle s_i \rangle \langle s_j \rangle \langle s_k \rangle
\end{aligned} \tag{9}$$

From the Legendre transformation we know the relationship between the higher order derivatives of F and those of G

$$\begin{aligned}
\frac{\partial^2 F}{\partial \theta_i \partial \theta_j} &= -\left(\frac{\partial^2 G}{\partial m_k \partial m_l} \right)^{-1}_{ij} \\
\frac{\partial^3 F}{\partial \theta_i \partial \theta_j \partial \theta_k} &= -\sum_{\alpha\beta\gamma} \frac{\partial^2 F}{\partial \theta_i \partial \theta_\alpha} \frac{\partial^2 F}{\partial \theta_j \partial \theta_\beta} \frac{\partial^2 F}{\partial \theta_k \partial \theta_\gamma} \frac{\partial^3 G}{\partial m_\alpha \partial m_\beta \partial m_\gamma}
\end{aligned} \tag{10}$$

If we make use of equation 8, with α set to one, we can estimate the higher order derivatives of G

$$\frac{\partial G}{\partial m_i} \approx \tanh^{-1} m_i + \frac{\partial \langle E_{int} \rangle}{\partial m_i} \tag{11}$$

$$\frac{\partial^2 G}{\partial m_i \partial m_j} \approx \frac{\delta_{ij}}{1 - m_i^2} + \frac{\partial^2 \langle E_{int} \rangle}{\partial m_i \partial m_j} \tag{12}$$

$$\frac{\partial^3 G}{\partial m_i \partial m_j \partial m_k} \approx \frac{2 \delta_{ijk} m_i}{(1 - m_i^2)^2} + \frac{\partial^3 \langle E_{int} \rangle}{\partial m_i \partial m_j \partial m_k} \tag{13}$$

where the derivatives of $\langle E_{int} \rangle$ are simple expressions in terms of m_i and the weights.

If we want to approximate the correlations given the thresholds and the weights, we first calculate the mean field variables m_i using equation 11 and then we use equation 12 to estimate the derivatives of G. After that we use the Legendre transformation to find the derivatives of F. Using equations 9 we can find the estimated correlations.

If we, however, know all the wanted correlations (as is the case if there are no hidden neurons) we can use equation 9 to calculate the derivatives of F and the Legendre transformation to obtain the derivatives of G. Since our approximated Gibbs free energy is linear in the weights, we can use equation 12 to make a direct estimate of the weights.

4 Results

We demonstrate the Linear Response approximation on a Boltzmann machine which has thresholds, second and third order weights. Our target distribution will be a Boltzmann distribution itself with only second and third order interactions

$$Q(\vec{s}) = \frac{1}{Z} \exp(\sum_{i<j} w_{ij} s_i s_j + \sum_{i<j<k} w_{ijk} s_i s_j s_k) \tag{14}$$

where Z is the normalization constant. The weights are random from a Gaussian with zero mean and a standard deviation $\sigma/\sqrt{N}$ for the second order and σ/N for the third order weights. N is the number of neurons. We have used a network with 10 neurons, since in that case it is possible to compare the results with the exact calculations.

First we have trained our Boltzmann machine using the exact learning algorithm and using the Linear Response approximation. We have learned the target distribution $Q(\vec{s})$ with a third order Boltzmann machine as well as with a second order Boltzmann machine (which we expect to give worse results). We use the Kullback divergence [4] to measure the distance between the target and the learned probability distribution.

In figure 1 the Kullback divergence is plotted versus the standard deviation σ. In the upper graph we used a target with third order interactions only; in the lower graph the target has both second and third order interactions. Since we have expanded G in the weights, we expect worse results as σ increases. Notice that the Kullback divergence of the exact third order machine is always zero since in that case the task is realizable.

From figure 1 we conclude that for targets in which third order correlations play a significant role the third order Linear Response approximation is useful as long as σ is not too large. (For a second order models it is known that the mean field approximation used here breaks down at $\sigma = 1/2$ for large N [5].)

Secondly we assess the quality of the approximation to compute the correlations of a given Boltzmann distribution. We have initialized the second and third order weights of a network of 10 neurons with zero mean and standard deviation $\sigma/\sqrt{N}$ for the second order interactions and σ/N for the third order interactions. We have plotted the exact correlations versus the estimated ones for a network with $\sigma = 0.25$. Figure 2 shows two graphs for $\langle s_i s_j \rangle$ and $\langle s_i s_j s_k \rangle$. We see that the approximated second and third order correlations are almost equal to the exact ones.

5 Discussion

In this paper we extended the Linear Response approximation to Boltzmann machines with higher order interactions. We showed that good approximations can be obtained as long as the weights in the target distribution are not too large.

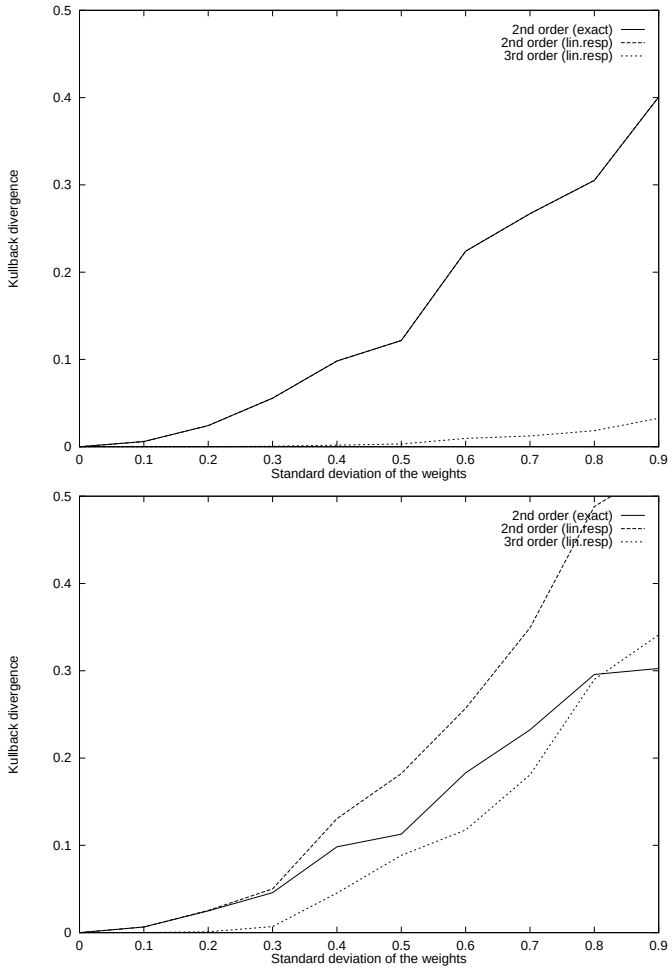

Figure 1: A network of 10 neurons learning a distribution a) with only third order interactions and b) with both second and third order interactions. Each point is an average over 10 random problems. Notice that in the upper graph the two second order lines are on top of each other.

It is possible to expand the Gibbs free energy G upto the second order of α. For a second order Boltzmann machine this brings in the TAP-term as was shown by Plefka [3] and Kappen et al. [2]. The same can be done for higher order Boltzmann machines which increases the accuracy of the estimation. The inversion of equation 12 however might be no longer possible since the derivatives of G are not linear in the weights in that case.

Although the theory presented is in principle valid for any order of the Boltzmann machine, only third and maybe fourth order will be useful in practice. The calculation time is polynomial, but it can be fairly large, since the

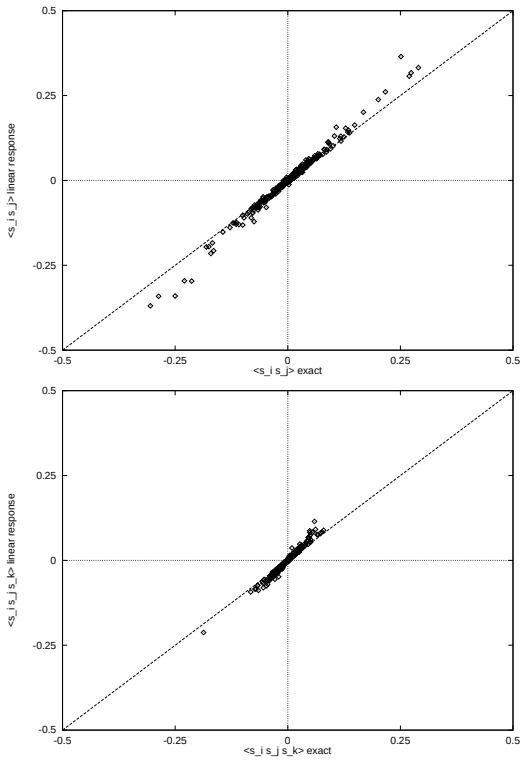

Figure 2: The estimation error of the second and third order correlations in a network of 10 neurons. a) $\langle s_i s_j \rangle_{ex}$ vs. $\langle s_i s_j \rangle_{lr}$ b) $\langle s_i s_j s_k \rangle_{ex}$ vs. $\langle s_i s_j s_k \rangle_{lr}$

exponent is proportional to the order of the Boltzmann machine. Furthermore the storage of all the correlations might be a problem.

References

[1] J. Hertz, A. Krogh, R.G. Palmer, Introduction to the Theory of Neural Computation, Chapter 7, Addison-Wesley (1991), ISBN 0-201-50395-6.

[2] H.J. Kappen and F.B. Rodríguez, Boltzmann Machine learning using Mean Field theory and Linear Response correction, Proceedings NIPS, MIT Press (1998). In press.

[3] T. Plefka, Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model, J. Phys. A: Math. Gen. 15 (1982) 1971–1978.

[4] S. Kullback, Information Theory and Statistics, Wiley, New York (1959).

[5] H. Takayama and K. Nemoto, Spin glass properties of a class of mean-field models, J. Phys.: Condens. Matter 2 (1990) 1997–2007.