
A mean field algorithm for Bayes learning in large feed-forward neural networks

Manfred Opper
Institut für Theoretische Physik
Julius-Maximilians-Universität, Am Hubland
D-97074 Würzburg, Germany
opper@physik.uni-wuerzburg.de

Ole Winther
CONNECT
The Niels Bohr Institute
Blegdamsvej 17
2100 Copenhagen, Denmark
winther@connect.nbi.dk

Abstract

We present a mean field algorithm to realise Bayes optimal predictions in large feed-forward networks. The derivation of the algorithm is based on methods developed within statistical mechanics of disordered systems. The algorithm also provides a leave-one-out cross-validation test of the predictions. Simulations show excellent agreement with theoretical results of statistical mechanics. Furthermore we derive other simpler mean field algorithms and make a comparison study.

Category: Theory, statistical mechanics
Oral presentation preferred
Correspondence should be addressed to Manfred Opper

1 INTRODUCTION

Bayes methods have become popular as a consistent framework for regularization and model selection in the field of neural networks (see e.g. [1]). In the Bayes approach to statistical inference [2] one assumes that the prior uncertainty about parameters of an unknown data generating mechanism can be encoded in a probability distribution, the so called *prior*. Using the prior and the likelihood of the data given the parameters, the *posterior* distribution of the parameters can be derived from Bayes rule. From this posterior, various estimates for functions of the parameter, like predictions about unseen data, can be calculated. However, in general, those predictions cannot be realised by specific parameter values, but only by an ensemble average over parameters according to the posterior probability.

Hence, exact implementations of Bayes method for neural networks require averages over network parameters which in general can be performed by time consuming Monte Carlo procedures. There are however useful approximate approaches for calculating posterior averages which are based on the assumption of a Gaussian form of the posterior distribution [1]. Under regularity conditions on the likelihood, this approximation becomes asymptotically exact when the number of data is large compared to the number of parameters. This Gaussian ansatz for the posterior may not be justified when the number of examples is small or comparable to the number of network weights. A second cause for its failure would be a situation where discrete classification labels are produced from a probability distribution which is a nonsmooth function of the parameters. This would include the case of a network with *threshold* units learning a *noise free* binary classification problem.

In this contribution we present an alternative approximate realization of Bayes method for neural networks, which is not based on asymptotic posterior normality. The posterior averages are performed using mean field techniques known from the statistical mechanics of disordered systems. Those are expected to become exact in the limit of a large number of network parameters under additional assumptions on the statistics of the input data. Our analysis follows the approach of Thouless, Anderson and Palmer (TAP) [3] as adapted to the simple perceptron by Mézard [8].

The basic set up of the Bayes method is as follows: We have a training set consisting of m input-output pairs $D_m = \{(\mathbf{s}^\mu, \sigma^\mu), m = 1, \dots, \mu\}$, where the outputs are generated independently from a conditional probability distribution $P(\sigma^\mu | \mathbf{w}, \mathbf{s}^\mu)$. This probability is assumed to describe the output σ^μ to an input \mathbf{s}^μ of a neural network with weights \mathbf{w} subject to a suitable noise process. If we assume that the unknown parameters \mathbf{w} are randomly distributed with a prior distribution $p(\mathbf{w})$, then according to Bayes theorem our knowledge about \mathbf{w} after seeing m examples is expressed through the posterior distribution

$$p(\mathbf{w} | D_m) = Z^{-1} p(\mathbf{w}) \prod_{\mu=1}^m P(\sigma^\mu | \mathbf{w}, \mathbf{s}^\mu) \quad (1)$$

where $Z = \int d\mathbf{w} p(\mathbf{w}) \prod_{\mu=1}^m P(\sigma^\mu | \mathbf{w}, \mathbf{s}^\mu)$ is called the partition function in statistical mechanics. Taking the average with respect to the posterior eq. (1), which in the following will be denoted by angle brackets, gives Bayes estimates for various quantities. For example the optimal predictive probability [5] for an output σ to a new input \mathbf{s} is given by $\hat{P}^{\text{Bayes}}(\sigma | \mathbf{s}) = \langle P(\sigma | \mathbf{w}, \mathbf{s}) \rangle$.

In section 2 exact equations for the posterior averaged weights $\langle \mathbf{w} \rangle$ for the simple perceptron are derived. Using a mean field ansatz these equations are solved in section 3. In section 4 we consider Bayes optimal predictions and a leave-one-out estimator. The approach is generalized in section 5 to a two-layer model, the tree committee machine [6]. Section 6 contains two different simplified mean field algorithms. We conclude in section 7 with a discussion of our results.

2 EXACT RESULTS FOR POSTERIOR AVERAGES

In this section we will derive equations for the posterior mean of the weights for the simple perceptron with N dimensional input, $\sigma(\mathbf{w}, \mathbf{s}) = \text{sign}(\Delta)$, where $\Delta = \frac{1}{\sqrt{N}} \mathbf{w} \cdot \mathbf{s}$ is the so called internal field. We will consider label noise in which the output is flipped, i.e. $\sigma \Delta < 0$ with a probability $(1 + e^\beta)^{-1}$. For simplicity, we will assume that the parameter β is known such that no prior on β is needed. The conditional

probability may thus be written as

$$P(\Delta^\mu) \equiv P(\sigma^\mu | \mathbf{w}, \mathbf{s}^\mu) = \frac{e^{-\beta \Theta(-\sigma^\mu \Delta^\mu)}}{1 + e^{-\beta}}, \quad (2)$$

where $\Theta(x) = 1$ for $x > 0$ and 0 otherwise. Obviously, this a nonsmooth function of the weights \mathbf{w} , for which the posterior will not become Gaussian asymptotically.

The average of the weights can be calculated from the distribution (1) by introducing the definition of Δ^μ and a conjugate parameter x^μ into the integrations by means of the δ -function

$1 = \int d\Delta^\mu \delta(\Delta^\mu - \frac{1}{\sqrt{N}} \mathbf{w} \cdot \mathbf{s}^\mu) = \int d\Delta^\mu \int_{-i\infty}^{+i\infty} \frac{dx^\mu}{2\pi i} e^{x^\mu (\frac{1}{\sqrt{N}} \mathbf{w} \cdot \mathbf{s}^\mu - \Delta^\mu)}$. We will consider the simplest case of a isotropic Gaussian prior $p(\mathbf{w}) = \frac{1}{\sqrt{2\pi}^N} e^{-\frac{1}{2} \mathbf{w} \cdot \mathbf{w}}$. In [7] anisotropic priors are treated as well. After straightforward Gaussian integrations we can show the following exact result

$$\langle \mathbf{w} \rangle = \frac{1}{\sqrt{N}} \sum_{\mu} \langle x^\mu \rangle \mathbf{s}^\mu \quad (3)$$

which states explicitly that the weights are linear combinations of the input vectors. This gives an example of the ability of Bayes method to regularize a network model: the effective number of parameters will never exceed the number of data points. To obtain the expectations over x^μ it is useful to introduce a reduced average for any

observable O , which is defined as $\langle O \rangle_{\mu} = \frac{\int d[\mathbf{w}] p(\mathbf{w}) O \prod_{\nu \neq \mu} P(\Delta^\nu)}{\int d[\mathbf{w}] p(\mathbf{w}) \prod_{\nu \neq \mu} P(\Delta^\nu)}$ over a posterior

where the μ -th pattern is kept out of the training set. Introducing an external field conjugate to x^μ we obtain after some calculations

$$\langle x^\mu \rangle = \langle \frac{\partial P(\Delta^\mu)}{\partial \Delta^\mu} \rangle_{\mu} / \langle P(\Delta^\mu) \rangle_{\mu} = \frac{\partial \ln \langle P(\Delta^\mu) \rangle_{\mu}}{\partial \langle \Delta^\mu \rangle_{\mu}} \quad (4)$$

which holds exactly for any prior. To express this average as a logarithmic derivative we have split the internal field into its average and fluctuating parts, i.e. we have set $\Delta^\mu = \langle \Delta^\mu \rangle_{\mu} + v^\mu$, with $v^\mu = \frac{1}{\sqrt{N}} (\mathbf{w} - \langle \mathbf{w} \rangle_{\mu}) \mathbf{s}^\mu$.

3 DERIVATION OF MEAN FIELD EQUATIONS

Sofar, no approximations have been made to obtain eqs. (3,4). However in order to end up with an explicit solution to equations (4) we will use mean field approximations based on the assumption of a central limit theorem for the fluctuating part of the internal field, v^μ which enters in the reduced average of eq. (4). We assume that the *non-Gaussian fluctuations* of w_i around $\langle w_i \rangle_{\mu}$ will sum up to make v^μ a Gaussian random variable. We expect that this gives a reasonable approximation, when N , the number of network weights is sufficiently large.¹ Following ideas of Mézard, Parisi and Virasoro [8, 4], who obtained mean field equations for a variety of disordered systems in statistical mechanics, one can argue that in many cases this assumption may be exactly fulfilled in the 'thermodynamic limit' $m, N \rightarrow \infty$ with $\alpha = \frac{m}{N}$ fixed. According to this ansatz, the reduced average becomes $\langle P(\Delta^\mu) \rangle_{\mu} = \int \frac{dv}{\sqrt{2\pi\lambda}} e^{-v^2/2\lambda} P(\langle \Delta_k \rangle_{\mu} + v)$, where λ^μ is the second moment

¹Note that the fluctuations of the internal field with respect to the *full* posterior mean (which depends on the input \mathbf{s}^μ) is *non* Gaussian, because the different terms in the sum become slightly correlated.

of v^μ which by definition is $\lambda^\mu \equiv \frac{1}{N} \sum_{i,j} s_i^\mu s_j^\mu (\langle w_i w_j \rangle_\mu - \langle w_i \rangle_\mu \langle w_j \rangle_\mu)$. In terms of λ^μ an explicit expression for the average in eq. (4) is now easily obtained

$$\langle P(\Delta^\mu) \rangle_\mu = \frac{1}{1 + e^{-\beta}} [e^{-\beta} + (1 - e^{-\beta}) H(-\sigma^\mu z^\mu)] , \quad (5)$$

where we used $e^{-\beta \Theta(-x)} = e^{-\beta} + (1 - e^{-\beta}) \Theta(x)$, $H(t) = \int_t^\infty dx e^{-x^2/2} / \sqrt{2\pi}$ and $z^\mu = \frac{\langle \Delta^\mu \rangle_\mu}{\sqrt{\lambda^\mu}} = \frac{\frac{1}{N} \langle \mathbf{w} \rangle \cdot \mathbf{s}^\mu - \lambda^\mu \langle x^\mu \rangle}{\sqrt{\lambda^\mu}}$. In the last step the reduced average $\langle \Delta_k^\mu \rangle_\mu = \langle \Delta_k^\mu \rangle - \langle v^\mu \rangle$ was rewritten in terms of the full posterior mean by noting that $\langle v^\mu \rangle = \langle P(\Delta^\mu) v^\mu \rangle_\mu / \langle P(\Delta^\mu) \rangle_\mu = \lambda^\mu \langle x^\mu \rangle$, using a standard theorem for Gaussian variables in the last equality. As an example consider the noiseless case $\beta \rightarrow \infty$ then eq. (4) becomes: $\langle x^\mu \rangle = \sigma^\mu \frac{e^{-z^{\mu 2}/2}}{\sqrt{2\pi \lambda^\mu H(-\sigma^\mu z^\mu)}}$.

The following approximation for λ^μ is expected to become exact in the thermodynamic limit if the inputs of the training set are drawn independently from a distribution, where all components s_i are uncorrelated and normalized i.e. $\overline{s_i} = 0$ and $\overline{s_i s_j} = \delta_{ij}$. The bars denote expectation over the distribution of inputs. For the generalisation to other input distribution see [7]. Our basic assumption is that the fluctuations of the λ^μ with the data set can be neglected so that we can replace them by their averages $\overline{\lambda^\mu}$. Since the reduced posterior averages are not correlated with the data s_i^μ , we obtain $\lambda^\mu \simeq \frac{1}{N} \sum_i (\langle w_i^2 \rangle_\mu - \langle w_i \rangle_\mu^2)$. Finally, we replace the reduced average by the expectation over the full posterior, neglecting terms of order $1/N$. Using $\sum_i \langle w_i^2 \rangle = N$, which follows from our choice of the Gaussian prior, we get $\lambda^\mu \simeq \lambda = 1 - \frac{1}{N} \sum_i \langle w_i \rangle^2$. This depends only on known quantities.

4 BAYES PREDICTIONS AND LEAVE-ONE-OUT

After solving the mean field equations by iteration we can make optimal Bayesian classifications for new data \mathbf{s} by choosing the output label with the largest predictive probability. In case of output noise this reduces to $\sigma^{\text{Bayes}}(\mathbf{s}) = \text{sign}(\langle \sigma(\mathbf{w}, \mathbf{s}) \rangle)$ Since the posterior distribution is independent of the new input vector we can apply the Gaussian assumption again to the internal field Δ of the new input and obtain $\sigma^{\text{Bayes}}(\mathbf{s}) = \sigma(\langle \mathbf{w} \rangle, \mathbf{s})$, i.e for the simple perceptron the averaged weights implement the Bayesian prediction. This will not be the case for multi-layer neural networks.

We can also get an estimate for the generalization error which occurs on the prediction of new data. The generalization error for the Bayes prediction for the perceptron is defined by $\epsilon^{\text{Bayes}} = \overline{\Theta(-\sigma(\mathbf{s}) \langle \sigma(\mathbf{w}, \mathbf{s}) \rangle)}$, where $\sigma(\mathbf{s})$ is the true output and the bar denotes average over the input distribution. To obtain the *leave-one-out estimator* of ϵ one removes the μ -th example from the training set and trains the network using only the remaining $m - 1$ examples. The μ 'th example is used for testing. Repeating this procedure for all μ an unbiased estimate for the Bayes generalization error with $m - 1$ training data is obtained as the mean value $\epsilon_{\text{MC}}^{\text{Bayes}} = \frac{1}{m} \sum_\mu \Theta(-\sigma^\mu \langle \sigma(\mathbf{w}, \mathbf{s}^\mu) \rangle_\mu)$ which is exactly the type of reduced averages which are calculated within our approach.

5 THE TREE COMMITTEE MACHINE

In the following we will consider the *tree committee machine* [6], composed of an input layer with N input units and a second layer of K hidden units with non-overlapping receptive fields. The network is built out of K subperceptrons with weight vectors \mathbf{w}_k , each having N/K couplings w_{jk} . A hidden neuron k computes

an individual output $\sigma_k = \text{sign}(\Delta_k)$ to its inputs s_{jk} , $j = 1, \dots, \frac{N}{K}$, where the internal field is given by $\Delta_k = \sqrt{\frac{K}{N}} \mathbf{w}_k \cdot \mathbf{s}_k$. The output neuron returns the majority vote $\sigma(\mathbf{w}, \mathbf{s}) = \text{sign}(K^{-\frac{1}{2}} \sum_k \sigma_k)$ as the final output of the net.

The derivation of the mean field equations follows along the same lines as for the simple perceptron considered above. We find (for a derivation see [7])

$$\langle \mathbf{w}_k \rangle = \sqrt{\frac{K}{N}} \sum_{\mu} \frac{\partial \ln \langle P(\Delta^{\mu}) \rangle_{\mu}}{\partial \langle \Delta_k^{\mu} \rangle_{\mu}} s_k^{\mu}. \quad (6)$$

In the simplifying case of a large number of hidden units, $K \rightarrow \infty$, in which one may apply the central limit theorem to the output of the hidden layer [6]

$$\langle P(\Delta^{\mu}) \rangle_{\mu} = \frac{1}{1 + e^{-\beta}} \left[e^{-\beta} + (1 - e^{-\beta}) H \left(-\sigma^{\mu} \frac{\sum_k g(z_k^{\mu})}{\sqrt{K}(1 - \sum_k g(z_k^{\mu})^2/K)} \right) \right], \quad (7)$$

where $g(x) = 1 - 2H(x)$, $z^{\mu} = \frac{\frac{K}{\sqrt{N}} \langle \mathbf{w}_k \rangle \cdot \mathbf{s}_k^{\mu} - \lambda_k \langle x_k^{\mu} \rangle}{\sqrt{\lambda_k}}$ and $\lambda_k = 1 - \frac{K}{N} \langle \mathbf{w}_k \rangle \cdot \langle \mathbf{w}_k \rangle$. The Bayesian prediction on a new input \mathbf{s} for the committee machine becomes

$$\sigma^{\text{Bayes}}(\mathbf{s}) = \text{sign} \left(\sum_k g \left(\sqrt{\frac{K}{N}} \frac{\langle \mathbf{w}_k \rangle \cdot \mathbf{s}_k}{\sqrt{\lambda_k}} \right) \right) \quad (8)$$

showing that the original architecture cannot implement Bayesian predictions. The original sign-activation function has to be changed to the sigmoid function, g with slope given by the the variances, λ_k . Only in the limit $\alpha \rightarrow \infty$ where the variances are expected to vanish, will the original architecture be able to implement Bayesian predictions.

6 SIMPLIFIED VERSIONS OF THE MEAN FIELD ALGORITHM

The first possible simplification of eq. (5) is to neglect the difference between the full and reduced average of the internal fields, i.e setting z^{μ} equal to $z_2^{\mu} \equiv \frac{\langle \Delta^{\mu} \rangle}{\sqrt{\lambda}} = \frac{1}{\sqrt{N\lambda}} \langle \mathbf{w} \rangle \cdot \mathbf{s}^{\mu}$. The second possible simplification is to furthermore set the variance to one $z_3^{\mu} \equiv \frac{1}{\sqrt{N}} \langle \mathbf{w} \rangle \cdot \mathbf{s}^{\mu}$. It can be shown that the latter approximation corresponds to a more traditional mean field method [10] which can also be derived from a variational treatment to yield an exact lower bound to the partition function.

In figure 1(b) we have made a comparison of the three mean field algorithms for the case a noisy ($\beta = 2$) simple perceptron with $N = 30$ input units. Note, that especially for small datasets the results are most robust against these simplifications.

7 CONCLUSION

In this paper we have presented a mean field algorithm which is expected to implement a Bayesian optimal classification well in the limit of large networks. So far the algorithm has been tested on problems, where the data are produced by a teacher network which has the same architecture as the learning network. It is important to test the robustness of the algorithm against deviations of the basic assumptions under which the procedure becomes exact. In this paper we assumed that the inputs had uncorrelated components, which for real world data will typically not be

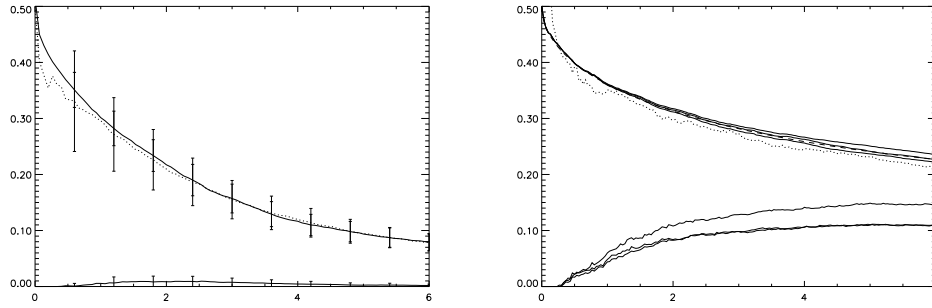


Figure 1: (a) Learning curve (generalization error ϵ versus $\alpha = m/N$) for the tree committee machine for $N = 60$ and $K = 3$ averaged over 200 runs. The output data are noise free and are produced from a teacher network of the same architecture. The full line is for Bayes learning (upper curve shows prediction error and the lower curve shows training error). The dotted line is the moving control estimate for the Bayes error. The moving control estimator has the larger error bars.

(b) Learning curves for the $N = 30$ simple perceptron with a noisy perceptron teacher network ($\beta = 2$), for the three mean field algorithms averaged over 100 runs. The dashed line is the theoretical prediction from [11]. The dotted line is the moving control estimator for the full mean field theory. The upper full lines is from bottom to top the results of the full mean field theory, the result from setting z^μ in eq. (5) equal to z_2^μ and z_3^μ (as defined in the text). The lower full lines is training error for the three algorithms in reverse order as the top lines. The error bars is of the same magnitude as in figure 1(a).

the case. How this will affect the reliability of the algorithm is an open question. From our simulations, we expect that the method is quite robust for small data sets which is in contrast to other approximations for Bayes method, which rely on expansions of the posterior for large data sets. It will be important to extend the algorithm to fully connected architectures. In that case it might be necessary to use the simplified versions of mean field method.

ACKNOWLEDGMENTS

This research is supported by a Heisenberg fellowship of the *Deutsche Forschungsgemeinschaft* and by the Danish Research Councils for the Natural and Technical Sciences through the Danish Computational Neural Network Center (CONNECT).

References

- [1] D. J. MacKay, *A practical Bayesian framework for backpropagation networks*, *Neural Comp.* **4** 448 (1992).
- [2] J. O. Berger, *Statistical Decision theory and Bayesian Analysis*, Springer-Verlag, New York (1985).
- [3] D.J. Thouless, P. W. Anderson, and R. G. Palmer, *Solution of 'Solvable model of a spin glass'* *Phil. Mag.* **35**, 593 (1977).
- [4] M. Mézard, *The space of interactions in neural networks: Gardner's calculation with the cavity method* *J. Phys. A* **22**, 2181 (1989).

- [5] V. N. Vapnik; *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, (1982).
- [6] H. Schwarze and J. Hertz, *Generalization in a Large Committee Machine*, Europhys. Lett. **20**, 375 (1992).
- [7] M. Opper and O. Winther, *A mean field approach to Bayes learning in feed-forward neural networks*, Phys. Rev. Lett. **76** 1964 (1996).
- [8] M. Mézard, G. Parisi and M. A. Virasoro. *Spin Glass Theory and Beyond*, Lecture Notes in Physics, 9, World Scientific, (1987).
- [9] M. Opper and D. Haussler, *Generalization Performance of Bayes Optimal Prediction Algorithm for Learning a Perceptron* Phys. Rev. Lett. **66**, 2677 (1991).
- [10] G. Parisi, *Statistical Field Theory*, Frontiers in Physics, Addison - Wesley, 1988.
- [11] M. Opper and D. Haussler, in *IVth Annual Workshop on Computational Learning Theory (COLT91)*, Morgan Kaufmann, 1991.