Mean Field Approach to Bayes Learning in Feed-Forward Neural Networks

Manfred Opper
Institut für Theoretische Physik, Julius-Maximilians-Universität, Am Hubland, D-97074 Würzburg, Germany

Ole Winther
CONNECT, The Niels Bohr Institute, Blegdamsvej 17, 2100 Copenhagen Ø, Denmark
(Received 6 October 1995)

We propose an algorithm to realize Bayes optimal predictions for feed-forward networks which is based on the Thouless-Anderson-Palmer mean field method developed for the statistical mechanics of disordered systems. We conjecture that our approach will be exact in the thermodynamic limit. The algorithm results in a simple built-in leave-one-out cross validation of the predictions. Simulations for the case of the simple perceptron and the committee machine are in excellent agreement with the results of replica theory.

PACS numbers: 87.10.+e, 64.60.Cn

Methods of statistical mechanics have been successfully applied to the understanding of a neural network's ability to infer an unknown rule from examples. For a review, see [1–3]. Based on the pioneering work of Gardner [4], the *typical* learning behavior of neural nets was studied by using statistical *ensembles* of networks. It was soon realized [5–7] that such an ensemble approach has a natural equivalent in the Bayesian methods of statistics. In the Bayesian approach [8] one assumes that the prior uncertainty about unknown parameters, e.g., network couplings, which have to be inferred from random data, can also be encoded in a probability distribution, the so called *prior*. On average the optimal prediction of novel data cannot be realized by a network in which the parameters take specific values, but only by performing an ensemble average over the so-called *posterior* distribution of parameters.

Several attempts have been made to implement Bayes methods as an algorithm in the context of neural network learning [7,9,10]. In this Letter we propose an algorithm for a multilayer network model that is expected to realize Bayes method exactly in the thermodynamic limit of large networks and large sample size, under the assumption that the input data are drawn at random from a known distribution. The basic idea is to perform averages over the posterior analytically, using mean field techniques well known in statistical mechanics. In contrast to the replica method [4], such a mean field calculation has to be performed for fixed random data. Our analysis follows the approach of Thouless, Anderson, and Palmer (TAP) [11] as adapted to the simple perceptron by Mézard [12], and provides a novel connection to Bayesian predictions.

To explain the Bayes method, let us assume that we observe a set of input-output data pairs $D_m = (\mathbf{s}^1, \sigma^1), \ldots, (\mathbf{s}^m, \sigma^m)$, where for fixed inputs the outputs σ^μ are generated independently from a conditional probability distribution $P(\sigma^\mu|\mathbf{w}, \mathbf{s}^\mu)$. In the context of neural networks, we can think of P as describing the output σ to an input $\mathbf{s}$ for a net with weights $\mathbf{w}$; the output may be corrupted by an independent noise process. If the unknown parameters $\mathbf{w}$ are randomly distributed with some prior probability $p(\mathbf{w})$, then according to Bayes theorem our knowledge about $\mathbf{w}$ after having observed m examples is expressed by the posterior density

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

where $Z = \int d[\mathbf{w}] p(\mathbf{w}) \prod_{\mu=1}^{m} P(\sigma^\mu|\mathbf{w}, \mathbf{s}^\mu)$ is the partition function. This equation enables us to calculate Bayes optimal estimates for various quantities. The best estimate for the weights $\hat{\mathbf{w}}$, which minimizes the cost function $\langle(\hat{\mathbf{w}} - \mathbf{w})^2\rangle$, is given by the posterior mean $\hat{\mathbf{w}} = \langle\mathbf{w}\rangle$. The angle brackets denote averages over the posterior (1). The optimal predictive probability [13] for an output σ to a new input $\mathbf{s}$ is given by $\hat{P}^{\text{Bayes}}(\sigma|\mathbf{s}) = \langle P(\sigma|\mathbf{w}, \mathbf{s})\rangle$. Our approach is based on using the TAP formalism to compute such averages.

In the following we will specialize on the *tree committee machine* [14], composed of an input layer with N input units and a second layer of K hidden units with nonoverlapping 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 $\sigma_k = \text{sgn}(\Delta_k)$ to its inputs s_{jk}, $j = 1, \ldots, N/K$, where the internal field is given by $\Delta_k = \sqrt{K/N}\, \mathbf{w}_k \cdot \mathbf{s}_k$. The output neuron returns the majority vote $\sigma(\mathbf{w}, \mathbf{s}) = \text{sgn}(K^{-1/2} \sum_k \sigma_k)$ as the final output of the net.

We will now derive equations for the posterior average of the weights $\langle w_{jk}\rangle$. The observation that the internal fields Δ_k become Gaussian in the thermodynamic limit $N \to \infty$ allows us to obtain closed equations for $\langle w_{jk}\rangle$ and to derive an expression for the Bayesian prediction $\hat{P}^{\text{Bayes}}(\sigma|\mathbf{s})$. We consider a simple noise model in which the output bit is flipped with probability $(1 + e^\beta)^{-1}$. Up

 0031-9007/96/76(11)/1964(4)$10.00

to a normalizing constant we may write $P(\sigma^\mu|\mathbf{w},\mathbf{s}^\mu)$ as

$$F(\Delta^\mu) \equiv \mathrm{Tr}_{\sigma_k^\mu=\pm1} \prod_k \Theta(\sigma_k^\mu \Delta_k^\mu) \times \left[e^{-\beta} + (1 - e^{-\beta}) \Theta\left(\sigma^\mu \frac{1}{\sqrt{K}} \sum_{i=1}^K \sigma_i^\mu \right) \right]. \tag{2}$$

We introduce auxiliary variables x_k^μ, and rewrite the partition function in (1) as

$$Z = \int d[\mathbf{w}] p(\mathbf{w}) \prod_\mu \left\{ \int \prod_k \left(\frac{dx_k^\mu \, d\Delta_k^\mu}{2\pi} \right) F(\Delta^\mu) \times \exp\left[i\sqrt{\frac{K}{N}} \sum_k x_k^\mu \mathbf{w}_k \cdot \mathbf{s}_k^\mu - i \sum_k x_k^\mu \Delta_k^\mu \right] \right\}. \tag{3}$$

We will consider a prior that factorizes over hidden units; the distribution for $\mathbf{w}_k$ is chosen to be Gaussian with covariance A_{ij}^k. By introducing fields conjugate to the weights, we can show that the posterior mean is given by

$$\langle w_{jk} \rangle = i\sqrt{\frac{K}{N}} \sum_\mu \langle x_k^\mu \rangle \sum_l A_{jl}^k s_{lk}^\mu . \tag{4}$$

In order to obtain the expectations of the variables x_k^μ it is useful to introduce the auxiliary average $\langle O \rangle_\mu = \int d[\mathbf{w}] \, p(\mathbf{w}) \, O \prod_{\nu\neq\mu} F(\Delta^\nu) / \int d[\mathbf{w}] \, p(\mathbf{w}) \prod_{\nu\neq\mu} F(\Delta^\nu)$ over a posterior, where the μth pattern is kept out of the training set. By introducing fields conjugate to x_k^μ, we obtain

$$i\langle x_k^\mu \rangle = \left\langle \frac{\partial F}{\partial \Delta_k^\mu} \right\rangle_\mu \Big/ \langle F \rangle_\mu = \frac{\partial \ln \langle F(\Delta^\mu) \rangle_\mu}{\partial \langle \Delta_k^\mu \rangle_\mu} \tag{5}$$

for an arbitrary prior. To express this average as a logarithmic derivative we split the internal field into its average and fluctuating parts, i.e., $\Delta_k^\mu = \langle \Delta_k^\mu \rangle_\mu + v_k^\mu$, with $v_k^\mu = \sqrt{K/N} \sum_j (w_{jk} - \langle w_{jk} \rangle_\mu) s_{jk}^\mu$ by definition.

While Eqs. (4) and (5) hold for arbitrary inputs and size of the network, the following mean field approximation is expected to be exact only in the thermodynamic limit $m, N \to \infty$, with $\alpha = m/N$ fixed. We have to assume that also the *inputs* are drawn independently from a distribution for which such a limit is meaningful.

We follow arguments based on the cavity approach of Parisi, Mézard, and Virasoro [12,15], who derived mean field equations of the TAP type for a variety of disordered systems. We assume that the non-Gaussian fluctuations of the w_{jk} with respect to $\langle w_{jk} \rangle_\mu$ sum up to a Gaussian distributed field v_k^μ in the thermodynamic limit. This assumption is valid if the phase space consists of only a single ergodic component. If there are many ergodic components, averages calculated in this way will correspond to a single component only. Note that the fluctuations of the internal field with respect to the *full* posterior mean (which depends on the input $\mathbf{s}^\mu$) is *non-Gaussian*, because the different terms in the sum become slightly correlated.

Using the Gaussian ansatz to compute the auxiliary averages, all we need to do in order to close our set of TAP equations (4) and (5) is an expression for the second moments of v_k^μ. Here we make the further assumption that these moments are self-averaging quantities when the inputs are independent random vectors with zero means and second moments $\overline{s_{ik} s_{jl}} = \delta_{kl} C_{ij}^k$, allowing for spatially organized data. In this case we get in the thermodynamic limit

$$\lambda_k \equiv \langle v_k^\mu v_n^\mu \rangle_\mu \simeq \delta_{kn} \frac{K}{N} \times \sum_{i,j} C_{ij}^k (\langle w_{ik} w_{jk} \rangle - \langle w_{ik} \rangle \langle w_{jk} \rangle) ; \tag{6}$$

in the last step we have replaced the auxiliary mean with the full mean neglecting terms of order $1/N$. Hence an expression for the yet unknown second moments of the weights $\mathbf{w}$ is now needed. In general, it can be also found from the cavity approach. Here we use a simpler approach, based on our choice of a Gaussian prior $p(\mathbf{w})$, the exact result $\sum_{ij} \langle w_{ik} w_{jk} \rangle (A^{-1})_{ij}^k = N/K$ follows then from (3). The choice $A^k = (C^k)^{-1}$ results in $\lambda_k = 1 - (K/N) \sum_{i,j} C_{ij}^k \langle w_{ik} \rangle \langle w_{jk} \rangle$.

An explicit expression for the average (5) is easily obtained using that the internal fields are Gaussian,

$$\langle F(\Delta^\mu) \rangle_\mu = \mathrm{Tr}_{\sigma_k^\mu=\pm1} \prod_k H(-\sigma_k^\mu z_k^\mu) \times \left[e^{-\beta} + (1 - e^{-\beta}) \Theta\left(\frac{1}{\sqrt{K}} \sum_{i=1}^K \sigma_i^\mu \right) \right], \tag{7}$$

where $H(t) = \int_t^\infty dx e^{-x^2/2} / \sqrt{2\pi}$ and

$$z_k^\mu = \frac{\langle \Delta_k^\mu \rangle_\mu}{\sqrt{\lambda_k}} = \frac{\sqrt{\frac{K}{N}} \langle \mathbf{w}_k \rangle \cdot \mathbf{s}^\mu - \lambda_k i \langle x_k^\mu \rangle}{\sqrt{\lambda_k}} . \tag{8}$$

The auxiliary average $\langle \Delta_k^\mu \rangle_\mu = \langle \Delta_k^\mu \rangle - \langle v_k^\mu \rangle$ is expressed in terms of the full one using a standard theorem for Gaussian random variables. Note that in this case the difference between the full and auxiliary mean is of order 1, and thus cannot be neglected.

By solving the set of nonlinear equations (4) and (5) with respect to the optimal coupling vectors $\langle \mathbf{w} \rangle$ we can make predictions on new data $\mathbf{s}$. One could directly implement the optimal vector $\langle \mathbf{w} \rangle$ into a committee machine and predict $\sigma^{\mathrm{opt}}(\mathbf{s}) = \sigma(\langle \mathbf{w} \rangle, \mathbf{s})$. A better approach is to use the optimal Bayes prediction for the output,

which in the case of output noise is given by $\sigma^{\text{Bayes}}(\mathbf{s}) = \text{sgn}\langle\sigma(\mathbf{w},\mathbf{s})\rangle$. Although this binary rule can no longer be implemented by a committee machine, it is possible to calculate σ^{Bayes} within our mean field approach. Since the posterior distribution is independent of the new input vector, we can again apply the Gaussian assumption to the internal field to obtain

$$\langle\sigma(\mathbf{w},\mathbf{s})\rangle = \text{Tr}_{\sigma_k=\pm1}\prod_k H\left(-\sigma_k\sqrt{\frac{K}{N}}\,\frac{\langle\mathbf{w}_k\rangle\cdot\mathbf{s}}{\sqrt{\lambda_k}}\right) \times \text{sgn}\left(K^{-1/2}\sum_i \sigma_i\right). \quad (9)$$

The Bayes prediction can be implemented by a neural network, using the architecture implied by (9). Obviously, $\sigma^{\text{Bayes}}(\mathbf{s})$ and $\sigma^{\text{opt}}(\mathbf{s})$ are different except for the simple perceptron ($K = 1$), a fact previously observed in [16].

Another important advantage of the mean field approach is the fact that by construction it yields an estimate for the generalization error which occurs on the prediction of new data. The generalization error for the Bayes prediction is defined by $\epsilon^{\text{Bayes}} = \langle\Theta(-\sigma(\mathbf{s})\langle\sigma(\mathbf{w},\mathbf{s})\rangle)\rangle_{\mathbf{s}}$, where $\sigma(\mathbf{s})$ is the output of the teacher network and $\langle\cdots\rangle_{\mathbf{s}}$ denotes average over the input distribution. To obtain the *moving control estimator* (or leave-one-out estimator) of ϵ one removes the μth example from the training set and trains the network using only the remaining $m-1$ examples. The μth example is used for testing. Repeating this procedure for $\mu = 1,\ldots,m$ and taking the mean value of the number of wrong classifications, we obtain an unbiased estimate for the Bayes generalization error with $m-1$ training data, which has the form $\epsilon^{\text{Bayes}}_{\text{MC}} = \frac{1}{m}\sum_\mu \Theta(-\sigma^\mu\langle\sigma(\mathbf{w},\mathbf{s}^\mu)\rangle_\mu)$. This expression fits nicely into the formalism of cavity fields and can be computed easily within the algorithm. The leave-one-out estimator $\epsilon^{\text{opt}}_{\text{MC}} = \frac{1}{m}\sum_\mu \Theta(-\sigma^\mu\sigma(\langle\mathbf{w}\rangle_\mu,\mathbf{s}^\mu))$ of the generalization error for the optimal learner $\epsilon^{\text{opt}} = \langle\Theta(-\sigma(\mathbf{s})\sigma(\langle\mathbf{w}\rangle,\mathbf{s}))\rangle_{\mathbf{s}}$ can be computed in a similar way.

We have tested the performance of the TAP mean field algorithm for the case of spherical inputs, $C_{ij} = \delta_{ij}$. In order to demonstrate that the mean field algorithm equations (4) and (5) can be solved and are able to generate the Bayes prediction we present simulations for the following two cases: (1) the noisy simple perceptron and (2) the noise-free committee machine with $K = 3$ hidden units. The simulation results are compared to theoretical results [17] and to the predictions of the moving control estimator. It turns out that the algorithm works very well even for small $N \approx 15$. In the simulations presented here, we set $N = 50$ for the simple perceptron and $N = 60$ for the committee machine. The $N + mK$ coupled equations (4) and (5) are solved by iteration from an initial solution, which is chosen to be the solution found for $m - X$ examples, with X between 1 and 10. The number of iterations required to reach the desired accuracy is typically 10–15. The simulation results are shown in Figs. 1 and 2. Figure 1 shows that there is very good agreement between theory and simulation, although the moving control estimate fluctuates very much for small training set sizes. For the committee machine the difference between ϵ^{Bayes} and ϵ^{opt} vanishes for large α. This can be understood from the fact that $\lambda_k \to 0$ for $\alpha \to \infty$, which implies $\sigma^{\text{opt}}(\mathbf{s}) \to \sigma^{\text{Bayes}}(\mathbf{s})$.

So far our approach has been tested on problems, where the output data were actually generated by a teacher net

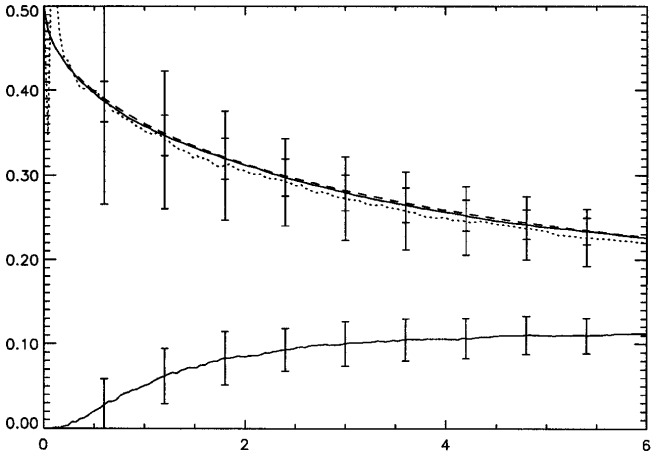

FIG. 1. The Bayes learning curve for the simple perceptron with output noise $\beta = 0.5$ and $N = 50$ averaged over 200 runs. The full lines are the simulation results (upper curve shows prediction error and the lower curve shows training error). The dashed line is the theoretical prediction. The dotted line with larger error bars is the moving control estimate.

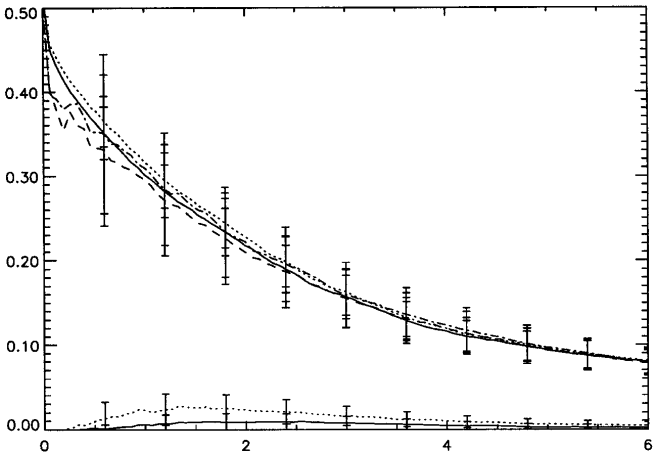

FIG. 2. The learning curve of the $K = 3$ committee machine with $N = 60$ averaged over 200 runs. The full line is for Bayes learning (upper curve shows prediction error and the lower curve shows training error). The dashed line is the moving control estimate for the Bayes error. The dotted line is for the optimal learner. The dash-dotted line is the moving control estimate for the optimal learner. The moving control estimators have the larger error bars.

of the same architecture as the student net. However, for mismatched architectures the weight space may be more complex as indicated by the occurrence of replica symmetry breaking [2] in the statistical mechanics studies of such models. In such cases, as is known for spin glass models [15,18], the possibility of having many solutions to the TAP equations may arise. This may deteriorate the convergence properties of the algorithm. In practical applications the data generating source is unknown; it might be possible to recognize that the proposed network model is a too bad approximation to the given data, through the inability to solve the TAP equations.

It is important to investigate the robustness of the TAP algorithm against violations of the basic theoretical assumptions. For instance, the input distribution is typically unknown for real world data, for which only empirical estimates of the correlations C_{ij} will be available. Extensive investigations of such questions and of that of extending the approach to fully connected architectures will be presented in forthcoming publications.

We thank Sara A. Solla for a critical reading of the manuscript. 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 Danisch Computational Neural Network Center (CONNECT).

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

[2] T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).

[3] M. Opper and W. Kinzel, in *Physics of Neural Networks*, edited by J. L. van Hemmen, E. Domany, and K. Schulten (Springer-Verlag, Berlin, 1995).

[4] E. Gardner, J. Phys. A **21**, 257 (1988).

[5] E. Levin, N. Tishby, and S. Solla, in *Proceedings of the 2nd Workshop on Computational Learning Theory* (Morgan Kaufmann, San Mateo, CA, 1989).

[6] G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses* (World Scientific, Singapore, 1990).

[7] M. Opper and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991).

[8] J. O. Berger, *Statistical Decision theory and Bayesian Analysis* (Springer-Verlag, New York, 1985).

[9] D. J. MacKay, Neural Comput. **4**, 448 (1992).

[10] T. L. H. Watkin, Europhys. Lett. **21**, 871 (1993).

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

[12] M. Mézard, J. Phys. A **22**, 2181 (1989).

[13] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data* (Springer-Verlag, Berlin, 1982).

[14] H. Schwarze and J. Hertz, Europhys. Lett. **20**, 375 (1992).

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

[16] O. Winther, B. Lautrup, and J-B. Zhang, Report No. CERN-TH/95-200 (to be published).

[17] M. Opper and D. Haussler, in *Proceedings of the IVth Annual Workshop on Computational Learning Theory (COLT91)* (Morgan Kaufmann, San Mateo, CA, 1991).

[18] A. J. Bray and M. A. Moore, J. Phys. C **13**, L469 (1980).