Baltzer Journals March 25, 1997

Query learning for maximum information gain in a multi-layer neural network

Peter Sollich*
Department of Physics, University of Edinburgh
Edinburgh EH9 3JZ, U.K.
E-mail: P.Sollich@ed.ac.uk

In supervised learning, the redundancy contained in random examples can be avoided by learning from queries, where training examples are chosen to be maximally informative. Using the tools of statistical mechanics, we analyse query learning in a simple multi-layer network, namely, a large tree-committee machine. The generalization error is found to decrease exponentially with the number of training examples, providing a significant improvement over the slow algebraic decay for random examples. Implications for the connection between information gain and generalization error in multi-layer networks are discussed, and a computationally cheap algorithm for constructing approximate maximum information gain queries is suggested and analysed.

1 Introduction

In supervised learning of input-output mappings, the traditional approach has been to study generalization from random examples. However, random examples contain redundant information, and generalization performance can thus be improved by *query learning*, where each new training input is selected on the basis of the existing training data to be most 'useful' in some specified sense. Query learning corresponds closely to the well-founded statistical technique of (sequential) *optimal experimental design*. In particular, we consider in this paper queries which maximize the expected information gain, which are related to the criterion of (Bayes) D-optimality in optimal experimental design. The generalization performance achieved by maximum information gain queries is by now well understood for single-layer neural networks such as linear and binary perceptrons [1, 2, 3]. For multi-layer networks, which are much more widely used in

*This work was partially supported by European Union grant no. ERB CHRX-CT92-0063.

practical applications, several heuristic algorithms for query learning have been proposed (see e.g., [4, 5]). While such heuristic approaches can demonstrate the power of query learning, they are hard to generalize to situations other than the ones for which they have been designed, and they cannot easily be compared with more traditional optimal experimental design methods. Furthermore, the existing analyses of such algorithms have been carried out within the framework of 'probably approximately correct' (PAC) learning, yielding worst case results which are not necessarily close to the potentially more relevant average case results. In this paper we therefore analyse the average generalization performance achieved by query learning in a multi-layer network, using the powerful tools of statistical mechanics. This is the first quantitative analysis of its kind that we are aware of.

2 The model

We focus our analysis on one of the simplest multi-layer networks, namely, the tree-committee machine (TCM). A TCM is a two-layer neural network with N input units, K hidden units and one output unit. The 'receptive fields' of the individual hidden units do not overlap, and each hidden units calculates the sign of a linear combination (with real coefficients) of the N/K input components to which it is connected. The output unit then calculates the sign of the sum of all the hidden unit outputs. A TCM therefore effectively has all the weights from the hidden to the output layer fixed to one. Formally, the output y for a given input vector $\mathbf{x}$ is

$$y = \operatorname{sgn}\left(\textstyle\sum_{i=1}^{K} \sigma_i\right) \qquad \sigma_i = \operatorname{sgn}\left(\mathbf{x}_i^{\mathrm{T}} \mathbf{w}_i\right) \tag{1}$$

where the σ_i are the outputs of the hidden units, $\mathbf{w}_i$ their weight vectors, and $\mathbf{x}^{\mathrm{T}} = (\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_K^{\mathrm{T}})$ with $\mathbf{x}_i$ containing the N/K (real-valued) inputs to which hidden unit i is connected. The N (real) components of the K (N/K)-dimensional hidden unit weight vectors $\mathbf{w}_i$, which we denote collectively by $\mathbf{w}$, form the adjustable parameters of a TCM. Without loss of generality, we assume the weight vectors to be normalized to $\mathbf{w}_i^2 = N/K$. We shall restrict our analysis to the case where both the input space dimension and the number of hidden units are large ($N \to \infty$, $K \to \infty$), assuming that each hidden unit is connected to a large number of inputs, i.e., $N/K \gg 1$. As our training algorithm we take (zero temperature) Gibbs learning, which generates at random any TCM (in the following referred to as a 'student') which predicts all the training outputs in a given set of p training examples $\Theta^{(p)} = \{(\mathbf{x}^\mu, y^\mu), \mu = 1 \ldots p\}$ correctly. We take the problem to be perfectly learnable, which means that the outputs y^μ corresponding to the inputs $\mathbf{x}^\mu$ are generated by a 'teacher' TCM with the same architecture as the student but with different, unknown weights $\mathbf{w}^0$. It is further assumed that there is no noise on the training examples. For learning from random examples, the training inputs $\mathbf{x}^\mu$ are sampled randomly from a distribution $P_0(\mathbf{x})$. Since the output (1) of a TCM is independent of the length of the hidden unit input vectors $\mathbf{x}_i$, we assume this distribution $P_0(\mathbf{x})$ to be uniform over all vectors $\mathbf{x}^{\mathrm{T}} = (\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_K^{\mathrm{T}})$ which obey the spherical constraints $\mathbf{x}_i^2 = N/K$. For query learning, the training inputs $\mathbf{x}^\mu$ are chosen to maximize the expected information gain of the student, as follows. The information gain is defined as the decrease in the entropy S in the parameter space of the student. The entropy for a training set $\Theta^{(p)}$ is given by

$$S(\Theta^{(p)}) = -\int d\mathbf{w} P(\mathbf{w}|\Theta^{(p)}) \ln P(\mathbf{w}|\Theta^{(p)}). \tag{2}$$

For the Gibbs learning algorithm considered here, $P(\mathbf{w}|\Theta^{(p)})$ is uniform on the 'version space', the space of all students which predict all training outputs correctly (and which satisfy the assumed spherical constraints on the weight vectors, $\mathbf{w}_i^2 = N/K$), and zero otherwise. Denoting the version space volume by $V(\Theta^{(p)})$, the entropy can thus simply be written as $S(\Theta^{(p)}) = \ln V(\Theta^{(p)})$. When a new training example $(\mathbf{x}^{p+1}, y^{p+1})$ is added to the existing training set, the information gain is $I = S(\Theta^{(p)}) - S(\Theta^{(p+1)})$. Since the new training output y^{p+1} is unknown, only the *expected* information gain, obtained by averaging over y^{p+1}, is available for selecting a maximally informative query $\mathbf{x}^{p+1}$. As derived in Ref. [2], the probability distribution of y^{p+1} given the input $\mathbf{x}^{p+1}$ and the existing training set $\Theta^{(p)}$ is $P(y^{p+1}\!=\!\pm1|\mathbf{x}^{p+1}, \Theta^{(p)}) = v^\pm$, where $v^\pm = V(\Theta^{(p+1)})|_{y^{p+1}=\pm1}/V(\Theta^{(p)})$. The expected information gain is therefore

$$\langle I \rangle_{P(y^{p+1}|\mathbf{x}^{p+1},\Theta^{(p)})} = -v^+ \ln v^+ - v^- \ln v^- \tag{3}$$

and attains its maximum value $\ln 2$ ($\equiv$ 1 bit) when $v^\pm = \frac{1}{2}$, i.e., when the new input $\mathbf{x}^{p+1}$ *bisects* the existing version space. This is intuitively reasonable, since $v^\pm = \frac{1}{2}$ corresponds to maximum uncertainty about the new output and hence to maximum information gain once this output is known.

Due to the complex geometry of the version space, the generation of queries which achieve exact bisection is in general computationally infeasible. The 'query by committee' algorithm proposed in Ref. [2] provides a solution to this problem by first sampling a 'committee' of $2k$ students from the Gibbs distribution $P(\mathbf{w}|\Theta^{(p)})$ and then using the fraction of committee members which predict +1 or -1 for the output y corresponding to an input $\mathbf{x}$ as an approximation to the true probability $P(y = \pm1|\mathbf{x}, \Theta^{(p)}) = v^\pm$. The condition $v^\pm = \frac{1}{2}$ is then approximated by the requirement that exactly k of the committee members predict +1 and -1, respectively. An approximate maximum information gain query can thus be found by sampling (or *filtering*) inputs from a stream of random inputs until this condition is met. The procedure is then repeated for each new query. As $k \to \infty$, this algorithm approaches the exact bisection algorithm, and it is on this limit that we focus in the following.

3 Exact maximum information gain queries

The main quantity of interest in our analysis is the generalization error $\epsilon_{\rm g}$, defined as the probability that a given student TCM will predict the output of the teacher incorrectly for a random test input sampled from $P_0(\mathbf{x})$. It can be expressed in terms of the overlaps $R_i = \frac{K}{N}\mathbf{w}_i^{\rm T}\mathbf{w}_i^0$ of the student and teacher hidden unit weight vectors $\mathbf{w}_i$ and $\mathbf{w}_i^0$ [6]. In the thermodynamic limit, the R_i are self-averaging, and can be obtained from a replica calculation of the average entropy S as a function of the normalized number of training examples, $\alpha = p/N$; details will be reported in a forthcoming publication [7]. The resulting average generalization error is plotted in Figure 1; for large α, one can show analytically that $\epsilon_{\rm g} \propto \exp(-\alpha\frac{1}{2}\ln 2)$. This exponential decay of the generalization error $\epsilon_{\rm g}$ with α provides a marked

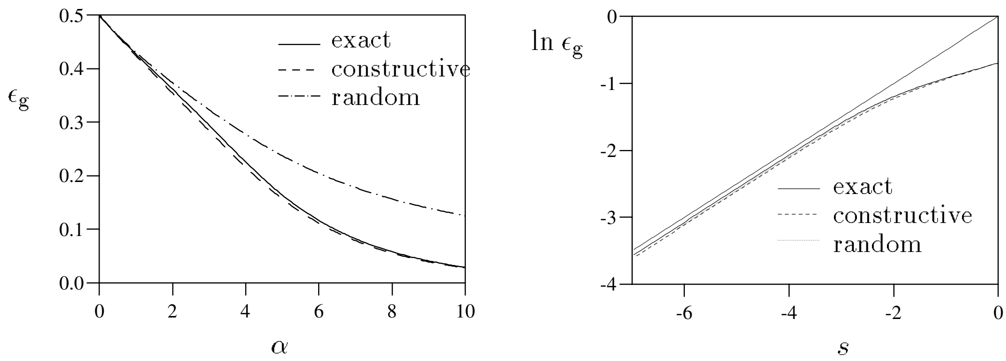

Figure 1: Left: Generalization error $\epsilon_{\rm g}$ vs. (normalized) number of examples α, for *exact* maximum information gain queries (Section 3), queries selected by *constructive* algorithm (Section 4), and *random* examples. Right: $\ln\epsilon_{\rm g}$ vs. (normalized) entropy s. For both queries and random examples, $\ln\epsilon_{\rm g} \approx \frac{1}{2}s$ (thin full line) for large negative values of s (corresponding to large α).

improvement over the $\epsilon_{\rm g} \propto 1/\alpha$ decay achieved by random examples [6]. The effect of maximum information gain queries is thus similar to what is observed for a binary perceptron learning from a binary perceptron teacher, but the decay constant c in $\epsilon_{\rm g} \propto \exp(-c\alpha)$ is only half of that for the binary perceptron [2]. This means that asymptotically, twice as many examples are needed for a TCM as for a binary perceptron to achieve the same generalization performance, in agreement with the results for random examples [6]. Since maximum information gain queries lead to an entropy $s = -\alpha\ln 2$ in both networks, we can also conclude that the relation $s \approx \ln\epsilon_{\rm g}$ for the binary perceptron [2] has to be replaced by $s \approx \ln\epsilon_{\rm g}^2$ for the tree committee machine. Figure 1 shows that, as expected, this relation holds independently of whether one is learning from queries or from random examples.

4 Constructive query selection algorithm

We now consider the practical realization of maximum information gain queries in the TCM. The query by committee approach, which in the limit $k \to \infty$ is an exact algorithm for selecting maximum information queries, filters queries from a stream of random inputs. This leads to an exponential increase of the query filtering time with the number of training examples that have already been learned [3]. As a cheap alternative we propose a simple algorithm for *constructing* queries, which is based on the assumption of an approximate decoupling of the entropies of the different hidden units, as follows. Each individual hidden unit of a TCM can be viewed as a binary perceptron. The distribution $P(\mathbf{w}_i|\Theta^{(p)})$ of its weight vector $\mathbf{w}_i$ given a set of training examples $\Theta^{(p)}$ has an entropy S_i associated with it, in analogy to the entropy (2) of the full weight distribution $P(\mathbf{w}|\Theta^{(p)})$. Our 'constructive algorithm' for selecting queries then consists in choosing, for each new query $\mathbf{x}^{\mu+1}$, the inputs $\mathbf{x}_i^{\mu+1}$ to the individual hidden units in such a way as to maximize the decrease in their entropies S_i. This can be achieved simply by choosing each $\mathbf{x}_i^{\mu+1}$ to be orthogonal to $\bar{\mathbf{w}}_i^{\mu} = \langle \mathbf{w}_i \rangle_{P(\mathbf{w}|\Theta^{(\mu)})}$ (and otherwise random, i.e., according to $P_0(\mathbf{x})$) [7], thus avoiding the cumbersome and time-consuming filtering from a random input stream. In practice, one would of course approximate $\bar{\mathbf{w}}_i^{\mu}$ by an average of $2k$ (say) samples from the Gibbs distribution $P(\mathbf{w}|\Theta^{(\mu)})$; these samples would have been needed anyway in the query by committee approach.

The generalization performance achieved by this constructive algorithm can again be calculated by the replica method; as shown in Figure 1, it is actually slightly superior to that of exact maximum information gain queries. The α-dependence of the entropy, $s = -\alpha \ln 2$, turns out to be the same as for maximum information gain queries; this indicates that the correlations between the individual hidden units become sufficiently small for $K \to \infty$, so that queries selected to minimize the individual hidden units' entropies also minimize the overall entropy of the TCM.

5 Conclusions

We have analysed query learning for maximum information gain in a large tree-committee machine (TCM). Or main result is the exponential decay of the generalization error ϵ_{g} with the normalized number of training examples α, which demonstrates that query learning *can* yield significant improvements over learning from random examples (for which $\epsilon_{\mathrm{g}} \propto 1/\alpha$ for large α) in multi-layer neural networks. The fact that the decay constant c in $\epsilon_{\mathrm{g}} \propto \exp(-c\alpha)$ differs from that calculated for single-layer nets such as the binary perceptron raises the question of how large c would be in more complex multi-layer networks. Combining the worst-case bound in [3] in terms of the VC-dimension with existing storage capacity bounds, one would estimate that c could be as small as $O(1/\ln K)$ for networks with a large number of hidden units K. This contrasts with our result $c \to$ const. for $K \to \infty$, and further work is clearly needed to establish whether there are realistic networks which saturate the lower bound $c = O(1/\ln K)$.

We have also analysed a computationally cheap algorithm for constructing (rather than filtering) approximate maximum information gain queries, and found that it actually achieves slightly better generalization performance than exact maximum information gain queries. This result is particularly encouraging considering the practical application of query learning in more complex multi-layer networks. For example, the proposed constructive algorithm can be modified for query learning in a fully-connected committee machine (where each hidden unit is connected to all the inputs), by simply choosing each new query to be orthogonal to the subspace spanned by the average weight vectors of *all* K hidden units. As long as K is much smaller than the input dimension N, and assuming that for large enough K the approximate decoupling of the hidden unit entropies still holds for fully connected networks, one would expect this algorithm to yield a good approximation to maximum information gain queries. The same conclusion may also hold for a *general* two-layer network with threshold units (where, in contrast to the committee machine, the hidden-to-output weights are free parameters), which can approximate a large class of input-output mappings. In summary, our results therefore suggest that the drastic improvements in generalization performance achieved by maximum information gain queries can be made available, in a computationally cheap manner, for realistic neural network learning problems.

References

[1] P. Sollich, Query construction, entropy, and generalization in neural network models, Phys. Rev. E, 49 (1994) 4637–4651.

[2] H. S. Seung, M. Opper, and H. Sompolinsky, Query by committee, in *Proc. 5th Workshop on Computational Learning Theory (COLT '92)*, ACM, New York, 1992, pp. 287–294.

[3] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, Information, prediction, and query by committee, in *Advances in Neural Information Processing Systems 5*, S. J. Hanson, J. D. Cowan, and C. Lee Giles eds., Morgan Kaufmann, San Mateo, CA, 1993, pp.483–490.

[4] E. Baum, Neural network algorithms that learn in polynomial time from examples and queries, IEEE Trans. Neural Netw., 2 (1991) 5–19.

[5] J.-N. Hwang, J. J. Choi, S. Oh, and R.J. Marks II, Query-based learning applied to partially trained multilayer perceptrons, IEEE Trans. Neural Netw., 2 (1991) 131–136.

[6] H. Schwarze and J. Hertz, Generalization in a large committee machine, Europhys. Lett., 20 (1992) 375–380.

[7] P. Sollich, Learning from minimum entropy queries in a large committee machine. In preparation.