Learning from minimum entropy queries in a large committee machine

Peter Sollich
Department of Physics, University of Edinburgh, Kings Buildings, Mayfield Road, Edinburgh EH9 3JZ, U.K.

In supervised learning, the redundancy contained in random examples can be avoided by learning from queries. Using statistical mechanics, we study learning from minimum entropy queries in a large tree-committee machine. The generalization error decreases exponentially with the number of training examples, providing a significant improvement over the algebraic decay for random examples. The connection between entropy and generalization error in multi-layer networks is discussed, and a computationally cheap algorithm for constructing queries is suggested and analysed.

To appear in **Physical Review E**, Rapid Communications. PACS numbers: 87.10.+e, 05.90.+m, 02.50.Wp.

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. In this paper, we consider *minimum entropy queries*, defined by maximizing the most common measure of 'usefulness', namely, the expected entropy decrease (or information gain). In order to achieve optimal generalization performance, the theoretically optimal choice of queries would of course be based on a direct minimization of the generalization error, and not on maximization of the entropy decrease. However, the generalization error is not in general accessible as an objective function for query selection, while the expected entropy decrease of a query can often be determined fairly easily. Since decrease in entropy and generalization error are normally correlated (see, e.g., Refs. [1,2]), minimizing entropy therefore provides a practical method for achieving near-optimal generalization performance by query learning.

The generalization performance achieved by minimum entropy queries is by now well understood for single-layer neural networks such as linear and binary perceptrons [1–3]. For multi-layer networks, which are much more widely used in practical applications, several heuristic algorithms for query learning have been proposed (see e.g., Refs. [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 techniques for query selection such as optimal experimental design. Furthermore, the existing analyses of such algorithms have been carried out within the framework of PAC (probably approximately correct) learning, yielding worst case bounds which do not necessarily represent average case behaviour. In this paper we therefore analyse the average generalization performance achieved by query learning in a multi-layer network, using the tools of statistical mechanics.

We focus on one of the simplest multi-layer neural networks, namely, the tree-committee machine (TCM). A TCM is a two-layer network with N input units, K hidden units and one output unit. The 'receptive fields' of the individual hidden units do not overlap, and all the weights from the hidden to the output layer are fixed to one. The output y for a given input vector $\mathbf{x}$ is therefore

$$y = \mathrm{sgn}\left(\frac{1}{\sqrt{K}}\sum_{i=1}^{K}\sigma_i\right) \qquad \sigma_i = \mathrm{sgn}\left(\sqrt{\frac{K}{N}}\mathbf{x}_i^{\mathrm{T}}\mathbf{w}_i\right) \quad (1)$$

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 which hidden unit i receives. The N 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, the weight vectors are assumed to be normalized to $\mathbf{w}_i^2 = N/K$, corresponding roughly to individual weights of $O(1)$.

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 decrease of the entropy S in the parameter space of the student. The entropy for a given training set $\Theta^{(p)}$ is defined as

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

For the Gibbs learning algorithm considered here, $P(\mathbf{w}|\Theta^{(p)})$ is uniform on the 'version space', the space of all students satisfying the spherical constraints $\mathbf{w}_i^2 = N/K$ which predict all training outputs correctly, and zero otherwise. Denoting the version space volume by $V(\Theta^{(p)})$, the entropy can thus simply be written as $S(\Theta^{(p)}) = \ln V(\Theta^{(p)})$. The entropy decrease $\Delta S = S(\Theta^{(p)}) - S(\Theta^{(p+1)})$ resulting from the addition of a new example $(\mathbf{x}^{p+1}, y^{p+1})$ to the existing training set cannot be maximized directly, since it depends on the new training output y^{p+1} generated by the unknown teacher. Queries are thus chosen to maximize the *expected* entropy decrease, obtained by averaging over y^{p+1}. Assuming a uniform prior over teachers, the probability of a certain teacher having produced the training set $\Theta^{(p)}$ is uniform over the version space and zero otherwise. The probability of obtaining output $y^{p+1} = \pm 1$ given input $\mathbf{x}^{p+1}$ is therefore simply $v^{\pm} = V(\Theta^{(p+1)})|_{y^{p+1}=\pm 1}/V(\Theta^{(p)})$, the fraction of the version space left over after the new example $(\mathbf{x}^{p+1}, y^{p+1} = \pm 1)$ has been added [3]. This gives the expected entropy decrease

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

which 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 [3] 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 $P(y = \pm 1|\mathbf{x}, \Theta^{(p)}) = v^{\pm}$. The condition $v^{\pm} = \frac{1}{2}$ is then approximated by the requirement that exactly k of the committee members predict output $+1$ and the other k predict -1 for the new training input $\mathbf{x}^{p+1}$. An approximate minimum entropy 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 exact bisection, and we focus on this limit in the following.

The main quantity of interest in our analysis is the generalization error ϵ_{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 $P_0(\mathbf{x})$. We consider the thermodynamic limit $N \to \infty$ at constant number of training examples per weight, $\alpha = p/N$, and focus on the case of a large number of hidden units, $K \to \infty$ with $N/K \gg 1$. The generalization error then takes the form [6]

$$\epsilon_{\mathrm{g}} = (1/\pi) \arccos R_{\mathrm{eff}} \qquad (3)$$

where R_{eff} is an effective overlap parameter given by

$$R_{\mathrm{eff}} = \frac{1}{K}\sum_{i=1}^{K} f(R_i) \qquad f(\cdot) = \frac{2}{\pi}\arcsin(\cdot)$$

in terms of the overlaps of the student and teacher hidden unit weight vectors, $R_i = (K/N)\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i^0$. In the thermodynamic limit, the R_i are self-averaging, i.e., their values for a specific teacher, training set and student from the Gibbs distribution are identical to their averages with probability one. These averages can be obtained from a replica calculation of the average entropy S as a function of α, following the calculations in Refs. [3,6]. We use the assumption of replica symmetry, which is believed to be exact for the case of noise free training data [6]. The replica calculation involves, in addition to the R_i, the overlap parameters ($\mu < p$)

$$q_i^p = (K/N)(\bar{\mathbf{w}}_i^p)^2 \qquad q_i^{\mu p} = (K/N)(\bar{\mathbf{w}}_i^p)^{\mathrm{T}}\bar{\mathbf{w}}_i^{\mu}$$

where $\bar{\mathbf{w}}_i^p = \langle \mathbf{w}_i \rangle_{P(\mathbf{w}|\Theta^{(p)})}$ and similarly for $\bar{\mathbf{w}}_i^{\mu}$. The $q_i^{\mu p}$ arise from the average over the $(\mu+1)$-th of the p training examples as the overlaps of the committee members which determine the selection of this example with the students trained on all p examples. The q_i^p can be determined from saddle point equations, whereas the $q_i^{\mu p}$ have to be determined independently. However, given the assumption of self-averaging of all overlap parameters, it can be shown that $q_i^{\mu p} = q_i^{\mu}$ in the case considered here [7]. This relation, which is proved by induction from the case $p = \mu + 1$, can be explained intuitively as follows. Given the first μ training examples $\Theta^{(\mu)}$, the teacher can be anywhere in the corresponding version space $\mathcal{V}^{\mu}$. Considering an average over all possible sets of training examples $\mu + 1 \ldots p$ produced by teachers in $\mathcal{V}^{\mu}$, the student is therefore equally likely to end up in any part of $\mathcal{V}^{\mu}$ after having been trained on the whole training set $\Theta^{(p)}$.

We assume symmetry between the hidden units [6], i.e., $q_i^p = q$, $q_i^{\mu p} = q_i^{\mu} = q(\alpha')$ ($\alpha' = \mu/N$) and $R_i = R$. The calculation, details of which will be reported elsewhere [7], can be further simplified by exploiting the relation $R = q$, which expresses the symmetry between teacher and student (see, e.g., Ref. [3]). One then obtains the normalized entropy $s = S/N$ (apart from an additive constant, which we fix such that $s = 0$ at $\alpha = 0$) as the saddle point of

$$\frac{1}{2}(q + \ln(1-q)) + 2\int_0^{\alpha} d\alpha' \int Dz\ H(\gamma z)\ln H(\gamma z) \quad (4)$$

with respect to q, where $\gamma = [(q_{\rm eff} - q_{\rm eff}(\alpha'))/(1-q_{\rm eff})]^{1/2}$ and $q_{\rm eff} = f(q)$, $q_{\rm eff}(\alpha') = f(q(\alpha'))$. We have also used the shorthand $Dz = dz \exp(-\frac{1}{2}z^2)/\sqrt{2\pi}$ and $H(z) = \int_z^\infty Dx$. Differentiating (4) with respect to α, one verifies that $ds/d\alpha = -\ln 2$ as expected for minimum entropy queries (the large committee limit $k \to \infty$ has already been taken) [8].

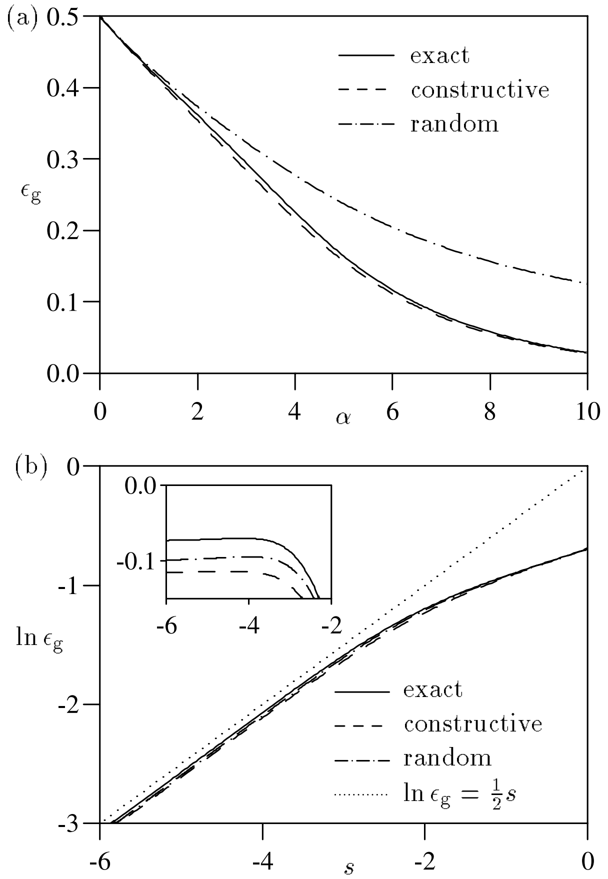

FIG. 1. (a) Generalization error $\epsilon_{\rm g}$ as a function of the normalized number of examples, α, for *exact* minimum entropy queries, queries as selected by *constructive* algorithm, and *random* examples. (b) Log generalization error $\ln \epsilon_{\rm g}$ vs. entropy s, for the same three cases. For both queries and random examples, $\ln \epsilon_{\rm g} \approx \frac{1}{2}s$ for large negative values of s (corresponding to large α). The very small separation between the curves is more clearly seen in the inset, which shows $\ln \epsilon_{\rm g} - \frac{1}{2}s$ vs. s.

Solving the saddle point equation numerically, we obtain the average generalization error as plotted in Figure 1(a). For large α, we find that $\epsilon_{\rm g} \propto \exp(-c\alpha)$ with $c = \frac{1}{2}\ln 2$, which can also be confirmed analytically from (4). This exponential decay of the generalization error $\epsilon_{\rm g}$ with α provides a marked improvement over the $\epsilon_{\rm g} \propto 1/\alpha$ decay achieved by random examples [6]. The effect of minimum entropy queries is thus similar to what is observed for a binary perceptron learning from a binary perceptron teacher, but the decay constant c is only half of that for the binary perceptron [3]. This means that asymptotically, twice as many examples are needed for a TCM as for a binary perceptron (when learning from a teacher with the respective architecture) to achieve the same generalization performance, in agreement with the corresponding result for random examples. Since in both networks, due to the binary nature of their outputs, minimum entropy queries lead to an entropy $s = -\alpha \ln 2$, we can also conclude that the large α relation $s \approx \ln \epsilon_{\rm g}$ for the binary perceptron [3] has to be replaced by $s \approx \ln \epsilon_{\rm g}^2$ for the tree committee machine. This relation should hold independently of whether one is learning from queries or from random examples. We have confirmed this by calculating the entropy for learning from random examples and comparing with the corresponding generalization error, as shown in Figure 1(b).

The above results are derived in the limit of a large number of hidden units, $K \to \infty$. For large but finite K they can be shown to be valid as long as the $O(1/K)$ correction to the generalization error (3), $(-1/2\pi K)R_{\rm eff}(1 - R_{\rm eff}^2)^{1/2}$, remains negligible, which holds for $\epsilon_{\rm g} \gg O(K^{-1/2})$. In the opposite regime $\epsilon_{\rm g} \ll O(K^{-1/2})$, i.e., for higher α, the generalization error $\epsilon_{\rm g} \approx (K/8)^{1/2}(1 - R_{\rm eff}) \propto \arccos(R)$ has the same functional dependence on R as for the binary perceptron, due to the fact that its dominant contribution arises from errors for which student and teacher only differ in the output of a single hidden unit. There is therefore a cross-over in the large α dependence of $\epsilon_{\rm g}$ from TCM ($K \to \infty$) to binary perceptron type behaviour around $\epsilon_{\rm g} = O(K^{-1/2})$.

We now consider the practical realization of minimum entropy queries in the TCM. The query by committee approach, which in the limit $k \to \infty$ is an exact algorithm for selecting minimum entropy 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 [1]. As a computationally 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 $\bar{\mathbf{w}}_i^\mu$ (and otherwise random, i.e., according to $P_0(\mathbf{x})$) [7], thus avoiding the 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.

An analysis of the generalization performance achieved by this constructive algorithm proceeds along the same line as the calculation for exact minimum entropy queries. Again restricting attention to the limit $k \to \infty$, we find that the saddle point expression (4) for the normalized entropy s still holds, but with γ now given by $\gamma = [a/(1-a)]^{1/2}$, $a = f([q - q(\alpha')]/[1 - q(\alpha')])$. Differentiating (4) with this replacement with respect to α, we find again that $ds/d\alpha = -\ln 2$, which means that in the thermodynamic limit that we consider, queries selected to minimize the individual hidden units' entropies also minimize the overall entropy of the TCM. This may seem surprising at first; heuristically, however, one can argue that for a large number of hidden units K, the correlations in the Gibbs distribution between the hidden unit weight vectors must be weak, and may indeed become negligible in the $K \to \infty$ limit considered here. The generalization performance achieved by the constructive query algorithm, shown in Figure 1(a), is actually slightly superior to that of exact minimum entropy queries as calculated in the previous section. This decrease in generalization error, although slight (about 4% for large α), exemplifies the fact that while decrease in entropy and in generalization error are normally correlated, there is no exact one-to-one relationship between them (compare the discussion in Ref. [2]). Query selection algorithms which achieve the same entropy decrease can therefore lead to different generalization performance.

We have found above a modification of the relationship between entropy s and generalization error $\epsilon_{\rm g}$ from $s \approx \ln \epsilon_{\rm g}$ for the binary perceptron to $s \approx \ln \epsilon_{\rm g}^2$ for the TCM, and a corresponding change of the decay constant c in the asymptotic behaviour of the generalization error $\epsilon_{\rm g} \propto \exp(-c\alpha)$. This leads to the interesting question of the value of c in more general multi-layer neural networks, and in particular its dependence on the number of hidden units K. The bound in Ref. [1], derived for the $k = 1$ query by committee algorithm, implies a lower bound on c which scales inversely with the VC-dimension [9] of the class of networks considered. Taking the storage capacity of a network as a coarse measure of its VC-dimension, one would then conclude from existing bounds [10] that c could be as small as $O(1/\ln K)$ for large K. However, the existing results for the capacity of particular networks like the TCM are not unambiguous enough to decide whether realistic networks would saturate this bound. Furthermore, it has been argued previously [11] that both the input space dimension *and* the VC-dimension determine the α-dependence of the generalization error. Replacing the VC-dimension in the bound in Ref. [1] with the input space dimension, one would then obtain a c of $O(1)$ independently of K. More theoretical work is clearly needed to clarify these questions.

With regard to the practical application of query learning in realistic multi-layer neural networks, the results we have obtained for a constructive query algorithm based on the assumption of a decoupling of the entropies of individual hidden units are encouraging. 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 minimum entropy queries [12]. It is an open question whether this conclusion would also hold for a general two-layer network with threshold units (where the hidden-to-output weights are also free parameters), which can approximate a large class of input-output mappings. We are currently investigating these issues in order to assess whether the significant improvements in generalization performance achieved by minimum entropy queries can be made available, in a computationally cheap manner, for learning in realistic binary output multi-layer neural networks.

---

[1] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, in *Advances in Neural Information Processing Systems 5*, edited by S. J. Hanson, J. D. Cowan, and C. L. Giles (Morgan Kaufmann, San Mateo, CA, 1993), pp. 483–490.

[2] P. Sollich, Phys. Rev. E **49**, 4637 (1994).

[3] H. S. Seung, M. Opper, and H. Sompolinsky, in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory (COLT '92)* (ACM, New York, 1992), pp. 287–294.

[4] E. Baum, IEEE Trans. Neural Netw. **2**, 5 (1991).

[5] J.-N. Hwang, J. J. Choi, S. Oh, and R. Marks II, IEEE Trans. Neural Netw. **2**, 131 (1991).

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

[7] P. Sollich (in preparation).

[8] The fact that the entropy is predicted correctly lends further support to the claim that the assumption of replica symmetry is correct for the problem considered here.

[9] V. Vapnik and A. Chervonenkis, Theory Probabil. Appl. **16**, 264 (1971).

[10] G. J. Mitchison and R. M. Durbin, Biological Cybernetics **60**, 345 (1989).

[11] M. Opper, Phys. Rev. Lett. **72**, 2113 (1994).

[12] To make the algorithm work once the permutation symmetry between hidden units is broken, one would of course have to restrict all weight space averages to one of the $K!$ 'ergodic' weight space sectors.