Published in:
Third European Symposium on Artificial Neural Networks (ESANN'95), Proceedings
Verleysen M (ed.), pages 217–222, D Facto, Brussels. 1995.

Minimum entropy queries for linear students learning nonlinear rules

Peter Sollich

Department of Physics, University of Edinburgh
Edinburgh EH9 3JZ, U.K.
P.Sollich@ed.ac.uk

Abstract. We study the fundamental question of how query learning performs in imperfectly learnable problems, where the student can only learn to approximate the teacher. Considering as a prototypical scenario a linear perceptron student learning a general nonlinear perceptron teacher, we find that queries for minimum entropy in student space (*i.e.,* maximum information gain) lead to the same improvement in generalization performance as for a noisy linear teacher. Qualitatively, the efficacy of query learning is thus determined by the structure of the student space alone; we speculate that this result holds more generally for minimum student space entropy queries in imperfectly learnable problems.

1. Introduction

The linear perceptron is arguably the simplest system that can learn from examples. It has recently been the subject of intensive investigation within the neural networks community (see, *e.g.,* [1, 2, 3, 4]). Traditionally, learning was assumed to be from a training set composed of *random examples*, with inputs chosen randomly and independently from some fixed distribution, and outputs provided by an unknown teacher, possibly corrupted by some noise. The aim is to generate, by a suitable training algorithm, a student linear perceptron which predicts as accurately as possible the outputs corresponding to inputs not contained in the training set, *i.e.,* which generalizes from the training data.

Since random training examples contribute less and less new information as the size of the training set grows, it is worthwhile investigating what improvements in generalization performance can be achieved by *learning from queries*, *i.e.,* by choosing each new training input such that it is, together with its corresponding output, most 'useful' in some specified sense. The most widely used measure of usefulness is the decrease of entropy, or gain of information, in the parameter space of the student (see, *e.g.,* [5]). The corresponding 'minimum entropy queries' have recently been studied for a *perfectly learnable* problem [6], where the teacher, like the student, is a linear perceptron. However, since in real-world problems the functional form of the teacher is almost never known, it is of fundamental importance to investigate the performance of query learning in *imperfectly learnable* problems, where the student can only learn to approximate the teacher. This we do in the present paper by considering as a prototypical imperfectly learnable scenario a linear perceptron student learning to approximate a teacher perceptron with a general nonlinear output function. We focus on the effect of the teacher nonlinearity on the efficacy of query learning, *i.e.,* the ability of queries to reduce the generalization error when compared to training on random examples.

In Section 2., we set up a formal model of the learning scenario considered. The main result of the paper for the average generalization error is presented and interpreted in Section 3.; the corresponding result for the training error is also given. We conclude with a summary and discussion of our results.

2. The learning scenario

We denote students by $\mathcal{N}$ (for '**N**eural network') and teachers by $\mathcal{V}$. A student $\mathcal{N}$ is specified by an N-dimensional weight vector $\mathbf{w}_\mathcal{N} \in \mathcal{R}^N$ and calculates its output $y_\mathcal{N}$ for an input vector $\mathbf{x} \in \mathcal{R}^N$ according to

$$y_\mathcal{N} = \tfrac{1}{\sqrt{N}}\mathbf{x}^T\mathbf{w}_\mathcal{N}.$$

Teachers are similarly parametrized in terms of a weight vector $\mathbf{w}_\mathcal{V} \in \mathcal{R}^N$, but calculate their output $y_\mathcal{V}$ by passing the (scaled) scalar product of $\mathbf{x}$ with this weight vector through a general nonlinear output function. Allowing the teacher outputs to be corrupted by noise, we only specify the average output for a given input

$$\langle y_\mathcal{V}\rangle_{P(y_\mathcal{V}|\mathbf{x},\mathcal{V})} = \bar{g}\left(\tfrac{1}{\sqrt{N}}\mathbf{x}^T\mathbf{w}_\mathcal{V}\right) \tag{1}$$

where $\bar{g}(\cdot)$ is a 'noise-averaged' output function. Concerning the noise process corrupting the teacher outputs, we make the mild assumption that the variance of the fluctuations $\Delta y_\mathcal{V}$ of the teacher outputs $y_\mathcal{V}$ around their average values (1) can be written as a function $\Delta^2(\cdot)$ of $\frac{1}{\sqrt{N}}\mathbf{x}^T\mathbf{w}_\mathcal{V}$ alone:

$$\langle(\Delta y_\mathcal{V})^2\rangle_{P(y_\mathcal{V}|\mathbf{x},\mathcal{V})} = \Delta^2\left(\tfrac{1}{\sqrt{N}}\mathbf{x}^T\mathbf{w}_\mathcal{V}\right).$$

This condition is fulfilled, for example, for additive noise with finite variance on the outputs or when the components of the teacher weight vector are corrupted by additive Gaussian noise with identical variance for each of the components.

We assume that the inputs are drawn from a uniform spherical distribution, $P(\mathbf{x}) \propto \delta(\mathbf{x}^2 - N\sigma_x^2)$. Using as our error measure the standard squared output deviation, $\frac{1}{2}(y_\mathcal{N} - y_\mathcal{V})^2$, we obtain for the generalization error, *i.e.,* the average error that a student $\mathcal{N}$ makes on an random test input when trying to approximate teacher $\mathcal{V}$,

$$\epsilon_\mathrm{g}(\mathcal{N},\mathcal{V}) = \frac{1}{2}\left[Q_\mathcal{N}\sigma_x^2 - 2\frac{R}{Q_\mathcal{V}}\langle h\bar{g}(h)\rangle_h + \langle\bar{g}^2(h)\rangle_h\right] + \frac{1}{2}\langle\Delta^2(h)\rangle_h \tag{2}$$

where

$$R = \tfrac{1}{N}\mathbf{w}_\mathcal{N}^T\mathbf{w}_\mathcal{V} \qquad Q_\mathcal{N} = \tfrac{1}{N}\mathbf{w}_\mathcal{N}^2 \qquad Q_\mathcal{V} = \tfrac{1}{N}\mathbf{w}_\mathcal{V}^2.$$

Here $\langle\cdot\rangle_h$ denotes an average over a Gaussian random variable h with zero mean and variance $Q_\mathcal{V}\sigma_x^2$, and we have assumed the 'thermodynamic limit', $N \to \infty$, of a perceptron with a very large number of input components.

As our training algorithm we take stochastic gradient descent on the training error, given by $E_\mathrm{t} = \frac{1}{2}\sum_\mu (y^\mu - y_\mathcal{N}(\mathbf{x}^\mu))^2$ for a training set consisting of p input-output pairs $\{(\mathbf{x}^\mu, y^\mu), \mu = 1\ldots p\}$. To prevent the student from overfitting noise in the data, we add a quadratic penalty term parametrized by a 'weight decay' parameter, λ, thus replacing E_t by $E = E_\mathrm{t} + \frac{1}{2}\lambda\sigma_x^2\mathbf{w}_\mathcal{N}^2$. Stochastic gradient descent on E leads to a Gibbs distribution of students, $P(\mathbf{w}_\mathcal{N}) \propto \exp(-E/T)$, where the 'learning temperature' T measures the amount of stochasticity in the training algorithm (see, *e.g.,* [1]). To have a well defined thermodynamic limit, we assume, as usual, that the number of training examples is proportional to the size of the perceptron, *i.e.,* $p = \alpha N$. We will concentrate our analysis on the average of the generalization error (2) over the post-training distribution of students, over all training sets produced by a given teacher $\mathcal{V}$, and over the prior distribution of teachers, which we assume to be Gaussian, $P(\mathcal{V}) \propto \exp(-\frac{1}{2}\mathbf{w}_\mathcal{V}^2/\sigma_\mathcal{V}^2)$.

For training on random examples, each input in the training set is drawn randomly and independently from the assumed uniform spherical input distribution. By contrast, for minimum entropy queries each new training input is chosen such that the entropy of the post-training distribution of students is minimized. For Gibbs learning, this entropy is given by (up to an irrelevant additive constant which depends on the learning temperature T only) [6]

$$S_\mathcal{N} = -\tfrac{1}{2}\ln\det\mathbf{M}_\mathcal{N} \qquad \mathbf{M}_\mathcal{N} = \lambda\sigma_x^2\mathbf{1} + \tfrac{1}{N}\textstyle\sum_\mu \mathbf{x}^\mu(\mathbf{x}^\mu)^T$$

where $\mathbf{1}$ denotes the $N \times N$ identity matrix. The independence of the entropy of the training outputs y^μ, and hence of the teacher output function $\bar{g}(\cdot)$, is characteristic of linear students. The entropy $S_\mathcal{N}$ is minimized by choosing each new training input along an eigendirection of the existing $\mathbf{M}_\mathcal{N}$ with minimal eigenvalue [6]. If we apply such minimum entropy queries in sequence, we find that the first N training inputs are pairwise orthogonal but otherwise random (on the sphere $\mathbf{x}^2 = N\sigma_x^2$), followed by another block of N such examples, and so on. It follows that the overlap $\frac{1}{N}(\mathbf{x}^\mu)^T\mathbf{x}^\nu$ of two different inputs for minimum entropy queries is smaller or equal to that for typical random inputs, which is $O(1/\sqrt{N})$. This simplifies the calculation by enabling us to expand averages like $\langle\bar{g}(\mathbf{w}_\mathcal{V}^T\mathbf{x}^\mu/\sqrt{N})\bar{g}(\mathbf{w}_\mathcal{V}^T\mathbf{x}^\nu/\sqrt{N})\rangle_{P(\mathcal{V})}$ in powers of $1/\sqrt{N}$, retaining, in the thermodynamic limit, only the lowest order terms.

3. Results

Following the calculation in [3] and using the techniques outlined in the previous section, we obtain as the main result of the paper the following expression for the average generalization error (primes denote derivatives):

$$\epsilon_\mathrm{g} = \frac{1}{2}\gamma_\mathrm{eff}^2\sigma_\mathcal{V}^2\sigma_x^2\left[\lambda_\mathrm{opt}G(\lambda) + \lambda(\lambda_\mathrm{opt} - \lambda)G'(\lambda)\right] + \epsilon_\mathrm{g,min}. \tag{3}$$

Here we have introduced the constants (h is again a zero mean Gaussian variable, now with variance $\sigma_{\mathcal{V}}^2\sigma_x^2$)

$$\begin{aligned} \gamma_{\rm eff} &= \langle h\bar{g}(h)\rangle_h/(\sigma_{\mathcal{V}}^2\sigma_x^2) = \langle \bar{g}'(h)\rangle_h \\ \sigma_{\rm act}^2 &= \langle \Delta^2(h)\rangle_h & \lambda_{\rm opt} &= \sigma_{\rm eff}^2/(\gamma_{\rm eff}^2\sigma_{\mathcal{V}}^2\sigma_x^2) \\ \sigma_{\rm eff}^2 &= \sigma_{\rm act}^2 + [\langle \bar{g}^2(h)\rangle_h - \langle h\bar{g}(h)\rangle_h^2/(\sigma_{\mathcal{V}}^2\sigma_x^2)] & \epsilon_{\rm g,min} &= \tfrac{1}{2}\sigma_{\rm eff}^2 \end{aligned} \tag{4}$$

The function G is the average of $\frac{\sigma_x^2}{N}\mathrm{tr}\,\mathbf{M}_{\mathcal{N}}^{-1}$ over the training inputs and is given by

$$G(\lambda) = \frac{1}{2\lambda}\left(1-\alpha-\lambda+\sqrt{(1-\alpha-\lambda)^2+4\lambda}\right) \tag{5}$$

for random examples [3], whereas for minimum entropy queries its value is [6]

$$G(\lambda) = \frac{\Delta\alpha}{\lambda+[\alpha]+1} + \frac{1-\Delta\alpha}{\lambda+[\alpha]} \tag{6}$$

where $[\alpha]$ is the greatest integer less than or equal to α and $\Delta\alpha = \alpha - [\alpha]$. In eq. (3) we have restricted ourselves to the case of zero learning temperature T, as finite T gives only an additional positive definite contribution $\frac{1}{2}TG(\lambda)$ to the average generalization error. For finite α, $\epsilon_{\rm g}$ is minimized when the weight decay parameter λ is set to its optimal value, $\lambda_{\rm opt}$; as $\alpha \to \infty$, the generalization error tends to its minimum achievable value, $\epsilon_{\rm g,min}$, which is independent of λ.

We now explain the remaining constants introduced in eqs. (4). The averages over h correspond to averages over the scalar product $\frac{1}{\sqrt{N}}\mathbf{x}^T\mathbf{w}_{\mathcal{V}}$. Therefore $\sigma_{\rm act}^2$ is the average variance of the teacher outputs, *i.e.,* the actual noise level. In order to clarify the meanings of $\gamma_{\rm eff}$ and $\sigma_{\rm eff}^2$, consider the special case of a linear teacher with 'gain constant' γ, given by $\bar{g}(h) = \gamma h$, and let the teacher outputs be corrupted by zero mean additive noise. It then follows that $\gamma_{\rm eff} = \gamma$ and $\sigma_{\rm eff}^2 = \sigma_{\rm act}^2$. The optimal weight decay $\lambda_{\rm opt} = \sigma_{\rm act}^2/\gamma^2\sigma_{\mathcal{V}}^2\sigma_x^2$ is the inverse of the mean-square signal-to-noise ratio of the teacher [6], and the minimum generalization error becomes $\epsilon_{\rm g,min} = \frac{1}{2}\sigma_{\rm act}^2$, which is simply the contribution from the noise on the teacher output. For a general nonlinear teacher and noise model, eqs. (4) can hence be interpreted as definitions of an appropriate effective gain constant and noise level, from which $\lambda_{\rm opt}$ and $\epsilon_{\rm g,min}$ are calculated just like for a linear teacher with additive output noise. For nonlinear $\bar{g}(\cdot)$, the (strictly positive) difference between $\sigma_{\rm eff}^2$ and $\sigma_{\rm act}^2$ represents effective noise arising from the fact that the linear student cannot reproduce the teacher perfectly. Due to this 'unlearnability' noise, the effective noise level $\sigma_{\rm eff}^2$ and the optimal weight decay $\lambda_{\rm opt}$ can be arbitrarily large even if there is no actual noise on the teacher outputs.

We have seen that the average generalization error obtained by learning to approximate a nonlinear teacher with a linear student is exactly the same as for an 'effective' noisy linear teacher. As a consequence, the efficacy of query learning is also the same as for a noisy linear teacher. Specifically, if we define

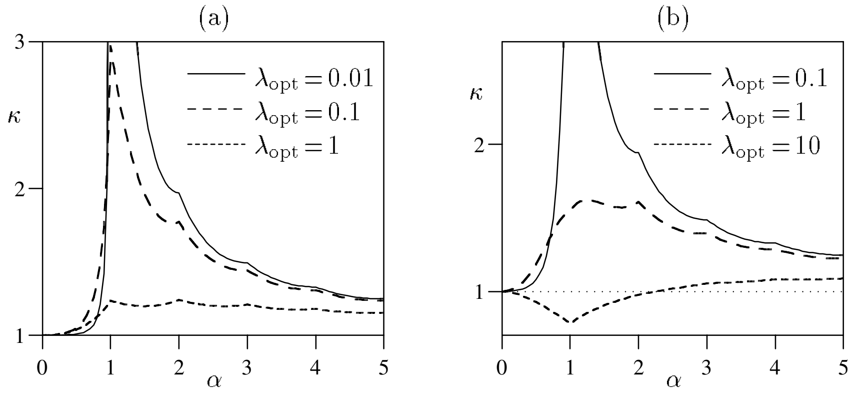

Figure 1: Relative improvement κ in generalization error due to minimum entropy queries, for (a) optimal weight decay, $\lambda = \lambda_{\rm opt}$, and (b) $\lambda = \lambda_{\rm opt}/10$.

the relative improvement in generalization performance due to querying, κ, as

$$\kappa(\alpha) = \frac{\epsilon_{\rm g}(\text{random examples}) - \epsilon_{\rm g,min}}{\epsilon_{\rm g}(\text{queries}) - \epsilon_{\rm g,min}}$$

then the result depends only on λ and $\lambda_{\rm opt}$, in the same way as for a noisy linear teacher. Figure 1 shows plots of $\kappa(\alpha)$ for some representative values of λ and $\lambda_{\rm opt}$. For large α, κ has the asymptotic expansion $\kappa = 1 + 1/\alpha + O(1/\alpha^2)$, which means that for $\alpha \to \infty$, random examples and queries yield the same generalization performance. This can be interpreted in the sense that for large α, learning is essentially hampered by (effective) noise in the data, for which queries are not much more effective than random examples (*cf.* the discussion in [6]). For finite α, the behaviour of κ depends on λ and $\lambda_{\rm opt}$. For optimal weight decay $\lambda = \lambda_{\rm opt}$ (Fig. 1a), κ has a maximum at $\alpha = 1$ whose height diverges as $1/\sqrt{\lambda_{\rm opt}}$ for $\lambda_{\rm opt} \to 0$; for $\lambda > \lambda_{\rm opt}$, the results are qualitatively similar. For $\lambda < \lambda_{\rm opt}$ (Fig. 1b), values of $\kappa < 1$ can occur which means that queries do *worse* than random examples. This case is particularly relevant for nonlinear teachers where $\lambda_{\rm opt}$ can be very large even if there is no actual noise on the teacher outputs. Nevertheless, the asymptotic expansion given above remains valid, and hence κ necessarily increases above one for large enough α.

We now briefly consider the training error in order to check whether it is affected by the teacher nonlinearity in the same way as the generalization error. To remove the trivial scaling with the number of training examples of the training error $E_{\rm t}$ introduced above, we consider the quantity $\epsilon_{\rm t} = E_{\rm t}/p$. Performing an average over students, training sets and teachers as before, and again restricting attention to the limit $T \to 0$, we find

$$\epsilon_{\rm t} = \epsilon_{\rm g,min} \left[1 - \frac{1}{\alpha} + \frac{\lambda^2}{\alpha\lambda_{\rm opt}} \Big(G(\lambda) + (\lambda - \lambda_{\rm opt}) G'(\lambda) \Big) \right]. \qquad (7)$$

The function $G(\lambda)$ is again given by eqs. (5,6) for random training examples and minimum entropy queries, respectively. We observe that the teacher nonlinearity only enters eq. (7) through $\epsilon_{\rm g,min}$ and $\lambda_{\rm opt}$, and hence affects the training error in exactly the same way as the generalization error. Note that for $\alpha \to \infty$, $\epsilon_{\rm t}$ tends to $\epsilon_{\rm g,min}$, as does the average generalization error $\epsilon_{\rm g}$. For random training examples, this is necessarily the case as the training error becomes an unbiased estimate of the generalization error for an infinite number of training examples. The fact that the result also holds for minimum entropy queries shows that they 'cover' the input space as well as random examples in the limit $\alpha \to \infty$; for queries chosen to optimize an objective function other than the student space entropy, this is not necessarily the case (*cf.* the discussion in [7]).

4. Summary and discussion

We have studied the performance of query learning in a prototypical imperfectly learnable scenario: a linear perceptron student learning a general nonlinear perceptron teacher. Our results show that for both the average generalization and training error, the effect of minimum entropy queries is the same for a nonlinear teacher as for a noisy linear teacher, with the noise level of this 'effective' linear teacher being the sum of the true noise level and an additional contribution arising from the fact that the problem is not perfectly learnable. Qualitatively, the improvement in generalization performance that can be obtained by query learning when compared to random examples is thus independent of whether the teacher rule that one is trying to learn is nonlinear or linear. We speculate that, in general, the qualitative effect of minimum student space entropy queries is determined by the structure of the student space and is essentially independent of the teacher space, *i.e.,* the class of rules that one is trying to approximate (see also [7]).

References

[1] A P Dunmur and D J Wallace. *J. Phys. A*, 26:5767–5779, 1993.

[2] A Bruce and D Saad. *J. Phys. A*, 27:3355–3363, 1994.

[3] A Krogh and J A Hertz. *J. Phys. A*, 25:1135–1147, 1992.

[4] S Bös, W Kinzel, and M Opper. *Phys. Rev. E*, 47:1384–1391, 1993.

[5] D J C MacKay. *Neural Comp.*, 4:590–604, 1992.

[6] P Sollich. *Phys. Rev. E*, 49:4637–4651, 1994.

[7] P Sollich and D Saad. Learning from queries for maximum information gain in imperfectly learnable problems. To be published in *Advances in Neural Information Processing Systems 7*, Morgan Kaufmann, San Francisco, 1995.