IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Learning unrealizable tasks from minimum entropy queries

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1995 J. Phys. A: Math. Gen. 28 6125

(http://iopscience.iop.org/0305-4470/28/21/016)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 171.66.16.68
The article was downloaded on 02/06/2010 at 02:15

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. 28 (1995) 6125–6142. Printed in the UK

Learning unrealizable tasks from minimum entropy queries

Peter Sollich[†]

Department of Physics, University of Edinburgh, Edinburgh EH9 3JZ, UK

Received 11 April 1995

Abstract. In supervised learning, learning from queries rather than from random examples can improve generalization performance significantly. We study the performance of query learning for unrealizable tasks, where the student cannot learn from the perfectly. As a simple model scenario of this kind, we consider a linear perceptron student learning a general nonlinear perceptron teacher. Two kinds of queries for maximum information gain, i.e. minimum entropy, are investigated: minimum *student space* entropy (MSSE) queries, which are appropriate if the teacher space is unknown, and minimum *teacher space* entropy (MTSE) queries, which can be used if the teacher space is assumed to be known, but a student of a simpler form has deliberately been chosen. We find that for MSSE queries, the structure of the student space determines the efficacy of query learning. MTSE queries, on the other hand, which we investigate for the extreme case of a binary perceptron teacher, lead to a higher generalization error than random examples, due to a lack of feedback about the progress of the student in the way queries are selected.

1. Introduction

In recent years, the powerful tools of statistical mechanics have been used successfully to study the behaviour of systems that learn from examples (for reviews see, for example, [1, 2].) The traditional approach has been to study generalization from random examples, where each example is an input–output pair with the input chosen randomly from some fixed distribution and the corresponding output provided by a teacher that the student is trying to approximate. However, the amount of novel information contained in random examples decreases towards zero as learning proceeds. Generalization performance can therefore be improved by learning from queries, i.e. by choosing the input of each new training example such that it will be, together with its expected output, in some sense 'maximally useful'. The most widely used measure of 'usefulness' is the information gain, i.e. the decrease in entropy of the post-training probability distributions in the parameter space of the student or the teacher.

We shall call the resulting queries 'minimum (student or teacher space) entropy (MSSE/MTSE) queries'; their effect on generalization performance has recently been investigated for *realizable tasks*, where student and teacher space are identical [3–5], and was found to depend qualitatively on the structure of the input–output mapping to be learned. For a linear perceptron, for example, one obtains a relative reduction in generalization error compared to learning from random examples which becomes insignificant as the number of training examples, p, tends to infinity. For a perceptron with binary output, on the other hand, minimum entropy queries result in a generalization error which decays exponentially

[†] E-mail address: P.Sollich@ed.ac.uk

0305-4470/95/216125+18$19.50 © 1995 IOP Publishing Ltd

as p increases—a marked improvement over the much slower algebraic decay with p in the case of random examples.

In practical situations, one very often encounters *unrealizable tasks*, where the student can only approximate the teacher, but not learn it perfectly. Unrealizable tasks can arise for two reasons. Firstly, the teacher space (i.e. the space of models generating the data) might be unknown. Because the teacher space entropy is then also unknown, MSSE (and not MTSE) queries have to be used for query learning. Secondly, the teacher space may be known, but a student of a simpler structure might have deliberately been chosen to facilitate or speed up training, for example. In this case, MTSE queries could be employed as an alternative to MSSE queries. The motivation for doing this would be strongest if, as in the learning scenario that we consider below, it is known from analyses of realizable tasks that the structure of the teacher space allows more significant improvements in generalization performance from query learning than the structure of the student space.

With the above motivation in mind, in this paper we investigate the performance of both MSSE and MTSE queries for a simple model unrealizable task, in which a linear perceptron student is trained on data generated by a general nonlinear perceptron teacher. Both student and teacher are specified by an N-dimensional weight vector with real components, and we will consider the thermodynamic limit $N \to \infty$, $p \to \infty$, with the normalized number of training examples, $\alpha = p/N =$ constant. Preliminary results for this scenario have been reported in [6].

Let us comment briefly on the practical relevance of the analysis of a learning scenario with a linear student. While it is true that in most applications of neural networks the nonlinearities present in standard feedforward networks play an important role, many fundamental insights into neural network learning have been obtained from analyses of linear model systems, where analytical solutions can be obtained [7–12]. Furthermore, it has been argued that the properties of networks with smooth nonlinearities can often be related to those of linear models by means of a local linearization procedure [13–15]. It is therefore reasonable to expect that at least qualitatively, the results of our analysis will to some extent carry over to more realistic feedforward neural networks.

The remainder of this paper is structured as follows. In section 2 below we formally define the learning scenario to be investigated. The generalization error for learning from random examples and from MSSE queries is calculated in section 3; MTSE queries are considered in section 4 for a binary perceptron teacher, which is in some way the most extreme case as explained below. We conclude in section 5 with a summary and discussion of our results.

2. The model

We denote students by $\mathcal{N}$ (for 'neural network') and teachers by $\mathcal{V}$ (for 'elements of the version space', see section 4). A student $\mathcal{N}$ is specified by an N-dimensional weight vector $w_{\mathcal{N}} \in \mathcal{R}^N$ and calculates its output $y_{\mathcal{N}}$ for an input vector $x \in \mathcal{R}^N$ according to

$$y_{\mathcal{N}} = \frac{1}{\sqrt{N}} x^T w_{\mathcal{N}} \, .$$

Teachers are similarly parametrized in terms of a weight vector $w_{\mathcal{V}} \in \mathcal{R}^N$, but calculate their output $y_{\mathcal{V}}$ by passing the (scaled) scalar product of x with this weight vector through a general nonlinear output function. Since we allow the teacher outputs to be corrupted by noise, we only specify the average output for a given input and assume that it can be written in the form

$$\langle y_{\mathcal{V}}\rangle_{P(y_{\mathcal{V}}|\boldsymbol{x},\mathcal{V})} = \bar{g}\left(\frac{1}{\sqrt{N}}\boldsymbol{x}^T\boldsymbol{w}_{\mathcal{V}}\right) \tag{2.1}$$

where $\bar{g}(\cdot)$ is a 'noise-averaged' output function. Implicit in equation (2.1) is the assumption that the noise process preserves, on average, the perceptron structure of the teacher. Similarly, we assume that the variance of the fluctuations $\Delta y_{\mathcal{V}}$ of the teacher outputs $y_{\mathcal{V}}$ around their average values (2.1) can be written as a function $\Delta^2(\cdot)$ of $\frac{1}{\sqrt{N}}\boldsymbol{x}^T\boldsymbol{w}_{\mathcal{V}}$ alone:

$$\langle(\Delta y_{\mathcal{V}})^2\rangle_{P(y_{\mathcal{V}}|\boldsymbol{x},\mathcal{V})} = \Delta^2\left(\frac{1}{\sqrt{N}}\boldsymbol{x}^T\boldsymbol{w}_{\mathcal{V}}\right). \tag{2.2}$$

This condition is fulfilled, for example, for additive noise on the outputs with finite, input-independent variance or (for inputs obeying a spherical constraint as considered below) when the components of the teacher weight vector are corrupted by additive Gaussian noise with identical variance for each of the components. Noise on the inputs, which has previously been studied with the aim of improving generalization performance (see, for example, [16–18]), can be treated similarly. For additive Gaussian noise on the input vector $\boldsymbol{x}$ (again with identical variance for each component), equation (2.2) holds as long as the length of *the teacher weight vector* $\boldsymbol{w}_{\mathcal{V}}$ *is fixed; this condition is enforced with probability one in the* thermodynamic limit for the Gaussian teacher prior considered below.

We assume that the inputs are drawn from a uniform spherical distribution, $P(\boldsymbol{x}) \propto \delta(\boldsymbol{x}^2 - N\sigma_x^2)$. 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 a random test input when trying to approximate teacher $\mathcal{V}$,

$$\begin{aligned}\epsilon_{\mathrm{g}}(\mathcal{N},\mathcal{V}) &= \tfrac{1}{2}\langle(y_{\mathcal{N}} - y_{\mathcal{V}})^2\rangle_{P(y_{\mathcal{V}}|\boldsymbol{x},\mathcal{V})P(\boldsymbol{x})}\\ &= \frac{1}{2}\left[Q_{\mathcal{N}}\sigma_x^2 + \langle\bar{g}^2(h)\rangle_h - 2\frac{R}{Q_{\mathcal{V}}}\langle h\bar{g}(h)\rangle_h\right] + \frac{1}{2}\langle\Delta^2(h)\rangle_h\end{aligned} \tag{2.3}$$

where

$$R = \frac{1}{N}\boldsymbol{w}_{\mathcal{N}}^T\boldsymbol{w}_{\mathcal{V}} \qquad Q_{\mathcal{N}} = \frac{1}{N}\boldsymbol{w}_{\mathcal{N}}^2 \qquad Q_{\mathcal{V}} = \frac{1}{N}\boldsymbol{w}_{\mathcal{V}}^2. \tag{2.4}$$

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. We have kept the last term in (2.3), which arises from the noise on the teacher outputs alone and could in principle be discarded, in order to make the comparison of linear and nonlinear teachers more transparent.

As our training algorithm we take stochastic gradient descent on the training error E_{t} which, for a training set $\Theta^{(p)} = \{(\boldsymbol{x}^\mu, y^\mu), \mu = 1\ldots p\}$, is

$$E_{\mathrm{t}} = \tfrac{1}{2}\sum_\mu (y^\mu - y_{\mathcal{N}}(\boldsymbol{x}^\mu))^2. \tag{2.5}$$

A weight decay term $\frac{1}{2}\lambda\sigma_x^2\boldsymbol{w}_{\mathcal{N}}^2$ is added for regularization, i.e. to prevent overfitting of noise in the training data, parametrized in terms of a dimensionless weight decay parameter λ. Stochastic gradient descent on the resulting energy function

$$E = E_{\mathrm{t}} + \tfrac{1}{2}\lambda\sigma_x^2\boldsymbol{w}_{\mathcal{N}}^2 \tag{2.6}$$

yields a Gibbs post-training distribution of students

$$P(\mathcal{N}|\Theta^{(p)}) \propto \exp(-E/T) \tag{2.7}$$

where the training temperature T measures the amount of stochasticity in the training algorithm. For the linear perceptron students considered here, this post-training distribution of students is a Gaussian distribution with mean $M_{\mathcal{N}}^{-1}a$ and covariance matrix $TM_{\mathcal{N}}^{-1}$, where (see, for example, [19] or any textbook on Bayesian statistics)

$$M_{\mathcal{N}} = \lambda\sigma_x^2\mathbf{1} + A \qquad A = \frac{1}{N}\sum_{\mu} x^{\mu}(x^{\mu})^T \qquad a = \frac{1}{\sqrt{N}}\sum_{\mu} y^{\mu}x^{\mu} \tag{2.8}$$

with $\mathbf{1}$ denoting the $N \times N$ unit matrix. To have a well defined thermodynamic limit, we assume, as usual, that $p = \alpha N$, i.e. that the number of training examples is proportional to the size of the perceptrons. We will concentrate our analysis on the average generalization error which can be obtained by successively averaging (2.3) over the post-training distribution of students, over the distribution of training sets $\Theta^{(p)}$ produced by a given teacher $\mathcal{V}$, and finally over the prior distribution of teachers, which we assume to be Gaussian, $P(\mathcal{V}) \propto \exp(-\frac{1}{2}w_{\mathcal{V}}^2/\sigma_{\mathcal{V}}^2)$. Under this prior, $Q_{\mathcal{V}} = \sigma_{\mathcal{V}}^2 + O(1/\sqrt{N})$, so that in the thermodynamic limit $Q_{\mathcal{V}}$ can be replaced by $\sigma_{\mathcal{V}}^2$ in (2.3). Hence the only non-trivial averages in the calculation of the average generalization error are the averages of the overlap parameters R and $Q_{\mathcal{N}}$ defined in (2.4). Note that typical deviations of the generalization error from its average value are $O(1/\sqrt{N})$ and are therefore vanishingly small in the thermodynamic limit.

The main aim of the present paper is to calculate, for the learning scenario defined above, the average generalization error as a function of the normalized number of training examples, $\alpha = p/N$, for learning from MSSE and MTSE queries, comparing the results to learning from random examples, and hence to draw conclusions about the efficacy of query learning in unrealizable tasks.

3. Random examples and minimum student space entropy (MSSE) queries

We now calculate the generalization performance resulting from random examples and from MSSE queries. For learning from random examples, each input in the training set is drawn randomly and independently from the assumed uniform spherical input distribution. By contrast, for MSSE queries each new training input is chosen such that the entropy of the post-training distribution of students is minimized. For the stochastic gradient descent learning algorithm described above and the resulting Gaussian post-training distribution, this entropy (normalized by N) is given by

$$S_{\mathcal{N}} = -\frac{1}{2N}\ln\det M_{\mathcal{N}} \tag{3.1}$$

up to an unimportant constant which depends on the learning temperature T only. The student space entropy is independent of the training outputs y^{μ}, which is characteristic of linear students (see, for example [20, 21]). The entropy (3.1) is minimized by choosing each new training input along an eigendirection of the existing $M_{\mathcal{N}}$ with minimal eigenvalue [5]. 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 $x^2 = N\sigma_x^2$), followed by another block of N such examples, and so on. The overlap $\frac{1}{N}(x^{\mu})^T x^{\nu}$ of two different inputs in a training set generated by MSSE queries is thus either 0 (if they belong to the same block) or of the size typical for random inputs, which is $O(1/\sqrt{N})$. These 'pseudo-random' overlaps simplify the calculation of the average generalization error, which is outlined in appendix A.

We obtain the following result for the average generalization error for learning from random examples and MSSE queries (primes denote derivatives):

$$\epsilon_{\mathrm{g}} = \tfrac{1}{2}\gamma_{\mathrm{eff}}^2\sigma_v^2\sigma_x^2\left[\lambda_{\mathrm{opt}}G(\lambda) + \lambda(\lambda_{\mathrm{opt}} - \lambda)G'(\lambda)\right] + \epsilon_{\mathrm{g,min}} . \tag{3.2}$$

Here we have introduced the constants

$$\gamma_{\mathrm{eff}} = \frac{1}{\sigma_v^2\sigma_x^2}\langle h\bar{g}(h)\rangle_h = \langle \bar{g}'(h)\rangle_h \tag{3.3}$$

$$\sigma_{\mathrm{eff}}^2 = \sigma_{\mathrm{act}}^2 + \langle \bar{g}^2(h)\rangle_h - \frac{1}{\sigma_v^2\sigma_x^2}\langle h\bar{g}(h)\rangle_h^2 \qquad \sigma_{\mathrm{act}}^2 = \langle \Delta^2(h)\rangle_h \tag{3.4}$$

$$\lambda_{\mathrm{opt}} = \frac{\sigma_{\mathrm{eff}}^2}{\gamma_{\mathrm{eff}}^2\sigma_v^2\sigma_x^2} \tag{3.5}$$

$$\epsilon_{\mathrm{g,min}} = \tfrac{1}{2}\sigma_{\mathrm{eff}}^2 \tag{3.6}$$

where $\langle\cdots\rangle_h$ denotes an average over a zero mean Gaussian random variable with variance $\sigma_v^2\sigma_x^2$. The function G is the average of $\frac{\sigma_x^2}{N}\operatorname{tr} M_N^{-1}$ over the training inputs and is given by

$$G(\lambda) = \frac{1}{2\lambda}\left(1 - \alpha - \lambda + \sqrt{(1 - \alpha - \lambda)^2 + 4\lambda}\right) \tag{3.7}$$

for random examples [7], whereas for MSSE queries its value is [5]

$$G(\lambda) = \frac{\Delta\alpha}{\lambda + \lfloor\alpha\rfloor + 1} + \frac{1 - \Delta\alpha}{\lambda + \lfloor\alpha\rfloor} \tag{3.8}$$

where $\lfloor\alpha\rfloor$ is the greatest integer less than or equal to α and $\Delta\alpha = \alpha - \lfloor\alpha\rfloor$. In equation (3.2) we have restricted ourselves to the case of zero learning temperature T as non-zero T gives only an additional positive-definite contribution $\frac{1}{2}TG(\lambda)$ to the average generalization error. For finite α, ϵ_{g} is minimized when the weight decay parameter λ is set to its optimal value, λ_{opt}, which is related to the effective signal-to-noise ratio of the teacher as explained below. As $\alpha \to \infty$, the generalization error tends to the minimum achievable value, $\epsilon_{\mathrm{g,min}}$, which is independent of λ as expected for the limit of an infinitely large training set.

We now explain the remaining constants introduced in (3.3)–(3.6). First note that, in all of the averages involved, $\sigma_v\sigma_x$ sets the scale of the arguments of $\bar{g}(\cdot)$ and $\Delta^2(\cdot)$. This was to be expected since, under the assumed input distribution and teacher space prior, $\frac{1}{\sqrt{N}}\boldsymbol{x}^T\boldsymbol{w}_v$ has zero mean and variance $\sigma_v^2\sigma_x^2$. In equation (3.4), σ_{act}^2 is the average variance of the fluctuations of the teacher outputs around their average, i.e. the actual noise level. In order to clarify the meaning of γ_{eff} and σ_{eff}^2, consider the special case of a linear teacher with 'gain constant' γ, which is given by $\bar{g}(h) = \gamma h$, and let the teacher outputs be corrupted by zero mean additive noise. It then follows that $\gamma_{\mathrm{eff}} = \gamma$ and $\sigma_{\mathrm{eff}}^2 = \sigma_{\mathrm{act}}^2$, and the minimum generalization error becomes $\epsilon_{\mathrm{g,min}} = \frac{1}{2}\sigma_{\mathrm{act}}^2$, which is simply the contribution from the noise on the teacher output. The optimal weight decay is $\lambda_{\mathrm{opt}} = \sigma_{\mathrm{act}}^2/\gamma^2\sigma_v^2\sigma_x^2$, which can be shown to be the inverse of the mean-square signal-to-noise ratio of the teacher [5]. For a general nonlinear teacher and noise model, we can interpret (3.3) and (3.4) as definitions of an appropriate effective gain constant and noise level, from which λ_{opt} and $\epsilon_{\mathrm{g,min}}$ are calculated just like for a linear teacher with additive output noise. The difference $\sigma_{\mathrm{eff}}^2 - \sigma_{\mathrm{act}}^2$ is greater than zero for nonlinear $\bar{g}(\cdot)$, and can be interpreted as effective noise arising from the fact that the linear student cannot reproduce the teacher perfectly. Note from (3.4) that this contribution to the effective noise can be very large for noise-averaged teacher output functions $\bar{g}(\cdot)$ containing a large part which is even in h. Since the effective gain γ_{eff} only depends on the odd part of $\bar{g}(\cdot)$, it follows from (3.5) that λ_{opt} can be arbitrarily large even if there is no actual noise on the teacher outputs.

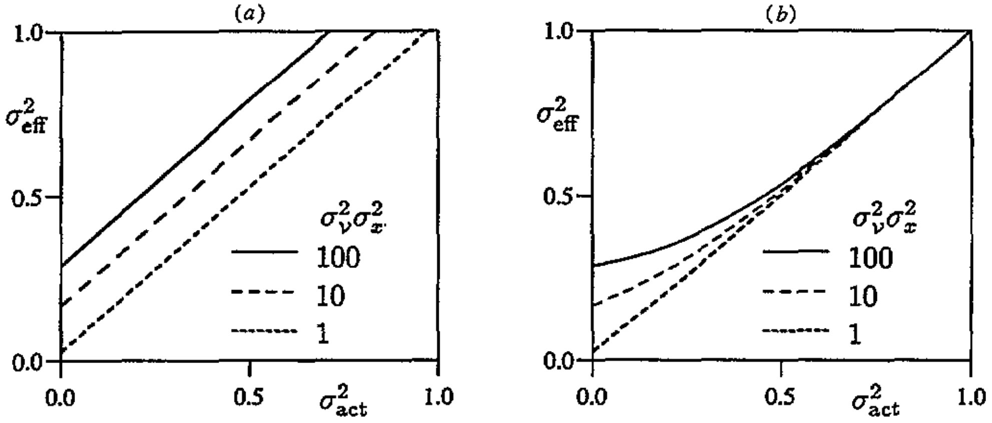

Figure 1. Effective noise level versus actual noise level for a teacher with tanh(·) output function, for additive Gaussian noise on (*a*) the outputs and (*b*) the components of the teacher weight vector. The curves are labelled by the values of $\sigma_V^2\sigma_x^2$.

By way of example, we show in figure 1 plots of $\sigma^2_{\rm eff}$ versus $\sigma^2_{\rm act}$ for a teacher with a tanh output function, for additive output noise (figure 1(*a*)), and for additive Gaussian noise with zero mean and identical variance on each of the N components of the teacher weight vector (figure 1(*b*)). In the latter case we have, denoting the noise variance by $\tilde{\sigma}_V^2$, $\tilde{g}(h) = \langle \tanh(h+\tilde{h})\rangle_{\tilde{h}}$ and $\Delta^2(h) = \langle \tanh^2(h+\tilde{h})\rangle_{\tilde{h}} - \tilde{g}^2(h)$, where $\tilde{h}$ is Gaussian with mean zero and variance $\tilde{\sigma}_V^2\sigma_x^2$. Applying equation (3.4) we obtain $\sigma^2_{\rm eff}$ and $\sigma^2_{\rm act}$ as functions of $\tilde{\sigma}_V$; eliminating $\tilde{\sigma}_V$ yields $\sigma^2_{\rm eff}$ as a function of $\sigma^2_{\rm act}$ as shown in figure 1(*b*). As $\sigma^2_{\rm act} \to 1$ (which corresponds to $\tilde{\sigma}_V \to \infty$), the difference $\sigma^2_{\rm eff} - \sigma^2_{\rm act}$ decreases towards zero because, with increasing $\tilde{\sigma}_V$, $\tilde{g}(\cdot)$ becomes approximately linear over an increasingly large range. Note that due to the nonlinearity of the teacher tanh(·) output function, $\sigma^2_{\rm eff}$ remains non-zero in all cases even for $\sigma^2_{\rm act} = 0$.

We have seen that the average generalization error obtained when learning to approximate a nonlinear teacher with a linear student is exactly the same as for a noisy linear teacher with an effective gain and noise level given by (3.3) and (3.4). Consequently, the efficacy of query learning for a nonlinear teacher is identical to that for a noisy linear teacher. Specifically, if we define 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 teacher nonlinearity enters the result only through the value of $\lambda_{\rm opt}$. Furthermore, the functional dependence on λ and $\lambda_{\rm opt}$ is the same as for a noisy linear teacher. Figure 2 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 [5]). For finite α, the behaviour of κ depends on λ and $\lambda_{\rm opt}$. For optimal weight decay $\lambda = \lambda_{\rm opt}$ (figure 2(*a*)), κ has a maximum at $\alpha = 1$ the height of which diverges as $\lambda_{\rm opt}^{-1/2}$ for $\lambda_{\rm opt} \to 0$. For $\lambda > \lambda_{\rm opt}$, the results are qualitatively similar but, for identical values of $\lambda_{\rm opt}$, κ is generally larger than for optimal weight decay $\lambda = \lambda_{\rm opt}$. For $\lambda < \lambda_{\rm opt}$ (figure 2(*b*)), κ

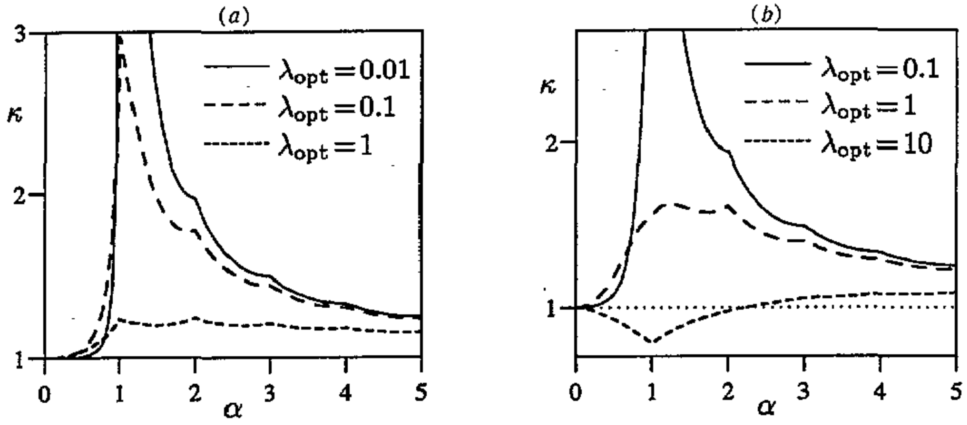

Figure 2. Relative improvement κ in generalization error due to MSSE queries, for (*a*) optimal weight decay, $\lambda = \lambda_{\rm opt}$, and (*b*) $\lambda = \lambda_{\rm opt}/10$.

tends to be smaller than for optimal weight decay; in fact, for $\lambda_{\rm opt} > 2$, values of $\kappa < 1$ can occur which means that queries do *worse* than random examples. As discussed in [5], this can be interpreted in the sense that for $\lambda < \lambda_{\rm opt}$, the weight decay 'underestimates' the effective teacher noise level, leading to spurious information gain in student space and thus making the student space entropy an unreliable indicator of generalization performance improvement. 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, even for $\lambda < \lambda_{\rm opt}$ the asymptotic expansion of $\kappa = 1 + 1/\alpha + O(1/\alpha^2)$ given above remains valid, and hence κ necessarily increases above one for large enough α.

The fact that κ tends to unity for $\alpha \to \infty$ implies that the relative improvement in generalization error over random examples due to MSSE querying tends to zero in this limit. In the next section we shall explore whether it is possible to improve generalization performance more significantly by using MTSE queries. Before doing so, however, we briefly mention the analogue of the result (3.2) for the average *training error*, in order to show that the training error is affected by the teacher nonlinearity in qualitatively the same way as the generalization error. To remove the trivial scaling with the number of training examples of the training error $E_{\rm t}$ defined in (2.5), we consider the quantity $\epsilon_{\rm t} = E_{\rm t}/p$. Performing an average over students, training sets and teachers as for the generalization error, we find

$$\epsilon_{\rm t} = \epsilon_{\rm g,min}\left[1 - \frac{1}{\alpha} + \frac{\lambda^2}{\alpha\lambda_{\rm opt}}\left(G(\lambda) + (\lambda - \lambda_{\rm opt})G'(\lambda)\right)\right]. \qquad (3.9)$$

As above, we have restricted ourselves to the case of zero training temperature T; non-zero T would give an additional positive contribution $T(1 - \lambda G(\lambda))/2\alpha$ to the average training error. The function $G(\lambda)$ is again given by (3.7) for random training examples and by (3.8) for MSSE queries. In equation (3.9) the teacher nonlinearity only enters through $\epsilon_{\rm g,min}$ and $\lambda_{\rm opt}$, and hence we again find the analogy between nonlinear and noisy linear teachers discussed above. Figure 3 shows plots of $\epsilon_{\rm t}(\alpha)$ for selected values of λ and $\lambda_{\rm opt}$. Interestingly, it can be shown that the training error is always smaller for MSSE queries than for random examples for $\lambda \leqslant \lambda_{\rm opt}$, whereas for $\lambda > \lambda_{\rm opt}$ it can also be greater. In comparison with the analogous relationships for the generalization error discussed above, the roles of the two λ-regimes are thus reversed here. For large α, the ratio of the training error for random examples to that for queries is $1+\lambda^2/(\lambda_{\rm opt}\alpha^3)+O(1/\alpha^4)$, which is always

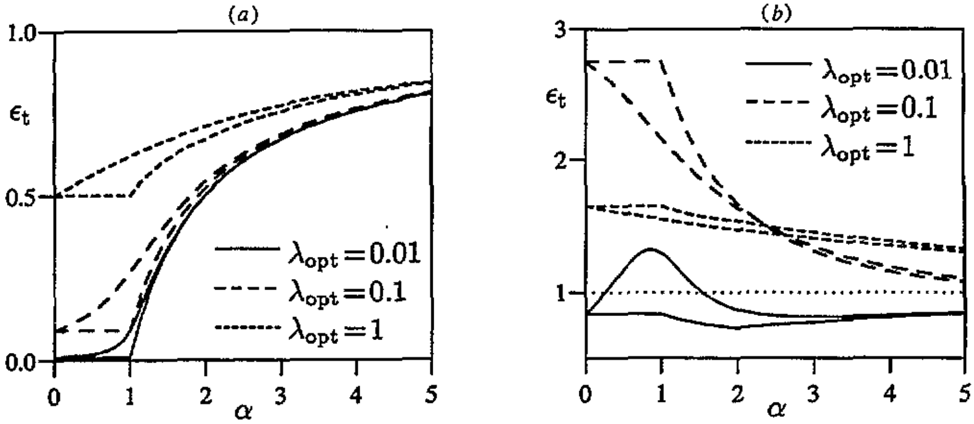

Figure 3. Average training error ϵ_t, in units of $\epsilon_{g,\min}$, for MSSE queries (curves which are constant for $\alpha \in [0, 1]$) and random examples. The weight decay parameter is (*a*) set to its optimal value, $\lambda = \lambda_{opt}$, and (*b*) $\lambda = 10\,\lambda_{opt}$.

larger than one for sufficiently large α.

Note that for $\alpha \to \infty$, ϵ_t tends to $\epsilon_{g,\min}$, as does the average generalization error ϵ_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 MSSE queries shows that they 'cover' the input space as well as random examples in the limit $\alpha \to \infty$. This is not necessarily the case for queries chosen to optimize an objective function other than the student space entropy. An example of this are the MTSE queries discussed in the next section, for which the generalization error tends to a limiting value for $\alpha \to \infty$ which depends on the weight decay λ, whereas the training error converges to $\frac{1}{2}$ in this limit, independently of λ, as shown in appendix B. In this case, therefore, the training error does not give an unbiased estimate of the generalization error, even for an infinite number of training examples.

4. Minimum teacher space entropy (MTSE) queries

We now consider the generalization performance achieved by MTSE queries. We remind readers that such queries could be employed if the teacher space is known, but a student of a simpler functional form has deliberately been chosen. As an example, consider a classification task, for which the teacher outputs are discrete class labels. In order to be able to use a training algorithm of gradient descent type, one might then choose to consider students with continuous outputs, for which the training error is a differentiable function of the student parameters. The scenario considered below, with a binary perceptron teacher and a linear perceptron student, can in fact be thought of as a simple model for situations of this kind. In general, the aim in using MTSE rather than MSSE queries would be to exploit the structure of the teacher space if this is known (for realizable tasks) to make query learning very efficient compared to random examples. In the binary teacher/linear student scenario, this is indeed the case: as mentioned in the introduction, the efficacy of minimum entropy query learning is high for a realizable task with binary perceptron student and teacher, whereas it is comparatively low when both student and teacher are linear perceptrons. In the unrealizable case, one would thus hope, by using MTSE queries, to 'transfer' the benefits for query learning of the binary perceptron structure of the teacher space into the student space.

The generalization performance achieved by MTSE and MSSE queries will differ most when the post-training student distribution and the posterior teacher distribution are maximally different. For continuous, invertible teacher output functions $\bar{g}(h)$, the posterior teacher distribution will be approximately Gaussian once the number of training examples is sufficiently large, and thus similar to the post-training student distribution (which, as explained above, is Gaussian for the linear students we are considering). This motivates our choice of considering a non-invertible teacher output function in our analysis of MTSE queries; specifically, we study the extreme case of an output function which only takes on two different values ±1, $\bar{g}(h) = \mathrm{sgn}(h)$, corresponding to a binary perceptron teacher. Since in this case the length of the teacher weight vector has no influence on the teacher's input–output mapping, we set $\sigma_{v}^{2} = 1$ without loss of generality. Similarly, the value of σ_{x}^{2} only scales the student overlap parameters R and $Q_{\mathcal{N}}$ and cancels from the average generalization error, and hence we also set $\sigma_{x}^{2} = 1$.

For simplicity, we assume that the training data generated by the binary perceptron teacher is noise free (corresponding to $\Delta^{2}(\cdot) \equiv 0$). The posterior probability distribution in teacher space given a certain training set is then proportional to the prior distribution on the *version space* (the set of all teachers that could have produced the training set without error) and zero everywhere else. From this the teacher space entropy (normalized by N) can be derived to be, up to an additive constant,

$$S_{v} = \frac{1}{N} \ln V$$

where the version space volume V is given by ($\Theta(z) = 1$ for $z > 0$ and 0 otherwise)

$$V = \int \mathrm{d}\boldsymbol{w}_{v}\, P(\boldsymbol{w}_{v}) \prod_{\mu=1}^{p} \Theta\left(y^{\mu} \frac{1}{\sqrt{N}} \boldsymbol{w}_{v}^{T} \boldsymbol{x}^{\mu}\right).$$

It can easily be verified that this entropy is minimized† by choosing queries $\boldsymbol{x}$ which 'bisect' the existing version space, i.e. for which the hyperplane perpendicular to $\boldsymbol{x}$ splits the version space into two equal halves [3, 4]. Such queries lead to an exponentially shrinking version space, $V(p) = 2^{-p}$, and hence a linear decrease of the entropy, $S_{v} = -\alpha \ln 2$. We consider instead queries which achieve qualitatively the same effect, but permit a much simpler analysis of the resulting student performance. They are similar to those studied in the context of a realizable task in [22], and are defined as follows. The $(p+1)$th query, $\boldsymbol{x}^{p+1}$, is obtained by first picking a random teacher vector $\overline{\boldsymbol{w}}_{v}$ from the version space defined by the existing p training examples, and then picking the new training input $\boldsymbol{x}^{p+1}$ from the distribution of random inputs but under the constraint that $\overline{\boldsymbol{w}}_{v}^{T} \boldsymbol{x}^{p+1} = 0$.

For the calculation of the student performance, i.e. the average generalization error, achieved by the approximate MTSE queries described above, we use an approximation based on the following observation. As the number of training examples, p, increases, the teacher vectors $\overline{\boldsymbol{w}}_{v}$ from the version space will align themselves with the true teacher $\boldsymbol{w}_{v}^{0}$; their components along the direction of $\boldsymbol{w}_{v}^{0}$ will increase, whereas their components perpendicular to $\boldsymbol{w}_{v}^{0}$ will decrease, varying widely across the $(N-1)$-dimensional hyperplane perpendicular to $\boldsymbol{w}_{v}^{0}$. Following [22], we therefore assume that the only significant effect of choosing queries $\boldsymbol{x}^{p+1}$ with $\overline{\boldsymbol{w}}_{v}^{T} \boldsymbol{x}^{p+1} = 0$ is on the distribution of the component of $\boldsymbol{x}^{p+1}$ along $\boldsymbol{w}_{v}^{0}$.

† More precisely, what is minimized is the value of the entropy after a new training example $(\boldsymbol{x}, y)$ is added, averaged over the distribution of the unknown new training output y given the existing training set and the new training input $\boldsymbol{x}$. See [5] for a more formal definition.

Writing this component as $x_0^{p+1} = (\boldsymbol{x}^{p+1})^T \boldsymbol{w}_\nu^0/|\boldsymbol{w}_\nu^0|$, its probability distribution can readily be shown to be

$$P(x_0^{p+1}) \propto \exp\left(-\tfrac{1}{2}\left(x_0^{p+1}/\sigma_x s_p\right)^2\right) \tag{4.1}$$

where s_p is the sine of the angle between $\overline{\boldsymbol{w}}_\nu$ and $\boldsymbol{w}_\nu^0$. For finite N, the value of s_p is dependent on the p previous training examples that define the existing version space and on the teacher vector $\overline{\boldsymbol{w}}_\nu$ sampled randomly from this version space. In the thermodynamic limit, however, the variations of s_p become vanishingly small and we can thus replace s_p by its average value, which is a function of p alone. In the thermodynamic limit, this average value becomes a continuous function of $\alpha = p/N$, the number of training examples per weight, which we denote simply by $s(\alpha)$. The calculation can then be split into two parts. First, the function $s(\alpha)$ is obtained from a calculation of the teacher space entropy using the replica method, generalizing the results of [23]. The average generalization error can then be calculated by using an extension of the response function method described in [24] or by another replica calculation (now in student space) as in [8]. Below, we only give the results of these calculations, deferring details to appendix B.

The first part of the calculation yields the teacher space entropy S_ν as the saddle point of

$$\frac{1}{2}\left(\frac{q-r^2}{1-q} + \ln(1-q)\right) + 2\int_0^\alpha \mathrm{d}\alpha' \int_0^\infty \mathrm{D}y \int_{-\infty}^\infty \mathrm{D}t \ln H\left(\frac{t\sqrt{q-r^2} - y\,r\,s(\alpha')}{\sqrt{1-q}}\right) \tag{4.2}$$

with respect to q and r, which are, respectively, the average scalar product (normalized by $N\sigma_\nu^2$) of two teachers from the version space, and of the true teacher and a teacher from the version space. Here we have used the abbreviations $\mathrm{D}z = \exp(-\frac{1}{2}z^2)\,\mathrm{d}z/\sqrt{2\pi}$ and $H(z) = \int_z^\infty \mathrm{D}z'$. The value of $s(\alpha)$ can be expressed in terms of the saddle-point value of r, which we denote by $r(\alpha)$, as $s^2(\alpha) = 1 - r^2(\alpha)$. The saddle-point equations yield $r(\alpha)$ and hence $s(\alpha)$ as a function of the values of $s(\alpha')$ for $0 \leqslant \alpha' < \alpha$. This determines the function $s(\alpha)$ recursively, starting from the initial condition $s(0) = 1$. Evaluating this recursion numerically, we obtain the results plotted in figure 4. For large α values, the teacher space entropy decreases linearly with α, with gradient $c \approx 0.44$, whereas the entropy for random examples, also shown for comparison, decreases much more slowly

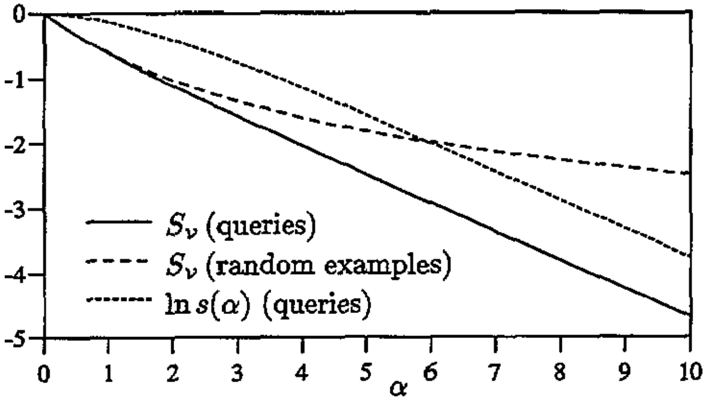

Figure 4. MTSE queries: teacher space entropy, S_ν (with value for random examples plotted for comparison), and $\ln s$, the log of the sine of the angle between the true teacher and a random teacher from the version space.

(asymptotically like $-\ln\alpha$ [23]). The linear α dependence of the entropy for queries corresponds to an average reduction of the version space volume with each new training example by a factor of $\exp(-c) \approx 0.64$, which is reasonably close to the factor $\frac{1}{2}$ for proper bisection of the version space. This shows that approximate MTSE achieve qualitatively the same results as true MTSE queries, and thus justifies our choice of analysing the former rather than the latter.

Before discussing the student performance achieved by (approximate) MTSE queries, we note from figure 4 that $\ln s(\alpha)$ decreases linearly with α for large α, with the same gradient as the teacher space entropy. Hence $s(\alpha) \propto \exp(-c\alpha)$ for large α, and MTSE queries force the teacher weight vectors from the version space to approach the true teacher exponentially quickly. It can easily be shown that if we were learning with a binary perceptron student, i.e. if we were considering a realizable task, then this would result in an exponentially decaying generalization error, $\epsilon_{\mathrm{g}} \propto \exp(-c\alpha)$. MTSE queries would thus lead to a marked improvement in generalization performance over random examples (for which $\epsilon_{\mathrm{g}} \propto 1/\alpha$ [23]). It is this significant benefit (in teacher space) of query learning that provides the motivation for using MTSE queries in unrealizable tasks such as the one considered here.

From the numerical values of $s(\alpha)$, the average generalization error achieved by the linear student when learning from our approximate MTSE queries can be calculated as outlined in appendix B. The results plotted in figure 5 show that MTSE queries do not have the desired effect of translating the benefits in teacher space into improvements in generalization performance for the linear student. In fact, they actually lead to a deterioration of generalization performance, i.e. a larger generalization error than that obtained for random examples. Worse still, they 'mislead' the student to such an extent that the minimum achievable generalization error is not reached even for an infinite number of training examples, $\alpha \to \infty$. How does this happen? It can be verified from (B.5) and (B.6) that the angle between the student and teacher weight vectors tends to zero for $\alpha \to \infty$ as expected, while $Q_{\mathcal{N}}$, the normalized squared length of the student weight vector, approaches

$$Q_{\mathcal{N}}(\alpha \to \infty) = \frac{2}{\pi}\left(\frac{\bar{s}(\infty)}{\lambda + \overline{s^2}(\infty)}\right)^2 \tag{4.3}$$

where $\bar{s}(\infty) = \int_0^\infty \mathrm{d}\alpha\, s(\alpha)$, $\overline{s^2}(\infty) = \int_0^\infty \mathrm{d}\alpha\, s^2(\alpha)$ as defined in (B.4). Unless the weight decay parameter λ happens to be equal to $\bar{s}(\infty) - \overline{s^2}(\infty)$, this is different from the optimal asymptotic value, which is $2/\pi$. This is the reason why, in general, the linear student does not reach the minimum possible generalization error even as $\alpha \to \infty$. The approach of $Q_{\mathcal{N}}$ to its non-optimal asymptotic value can cause an increase in the generalization error for large α and a corresponding minimum of the generalization error at some finite α, as can be seen in the plots for $\lambda = 0.01$ and 0.1 in figure 5. For $\lambda = 0$, equation (4.3) has the following intuitive interpretation. As α increases, the version space shrinks around the true teacher $\boldsymbol{w}_{\mathrm{v}}^{0}$, and hence MTSE queries become 'more and more orthogonal' to $\boldsymbol{w}_{\mathrm{v}}^{0}$. As a consequence, the distribution of training inputs along the direction of $\boldsymbol{w}_{\mathrm{v}}^{0}$ is narrowed down progressively (cf equation (4.1)). Trying to find a best fit to the teacher's binary output function over this narrower range of inputs, the linear student learns a function which is steeper than the best fit over the range of random inputs (which would give minimum generalization error). This corresponds to a suboptimally large length of the student weight vector, in agreement with (4.3): $Q_{\mathcal{N}}(\alpha \to \infty) > 2/\pi$ for $\lambda = 0$ because $\overline{s^2}(\infty) < \bar{s}(\infty)$.

Summarizing the results of this section, we have found that although MTSE queries are very beneficial in teacher space, they are entirely misleading for the linear student, to the extent that the student does not learn to approximate the teacher optimally even for

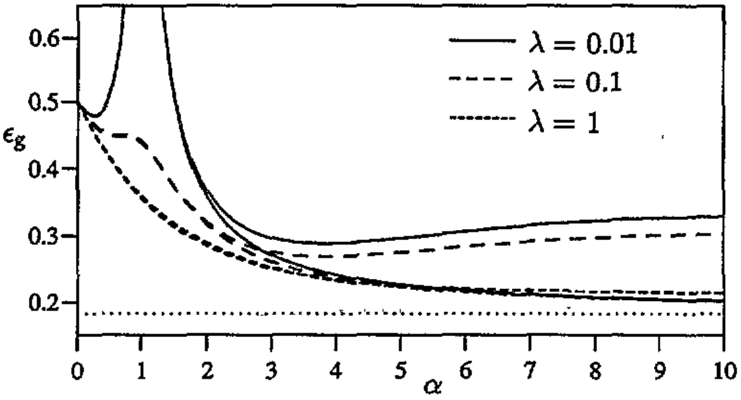

Figure 5. Generalization error for MTSE queries (higher curves of each pair) and random examples (lower curves), for weight decay $\lambda = 0.01, 0.1, 1$. The curves for random examples (which are virtually indistinguishable from one a nother already at $\alpha = 10$) converge to the minimum achievable generalization error $\epsilon_{g,\min}$ (short-broken curve) as $\alpha \to \infty$.

an infinite number of training examples. With the benefit of hindsight, we note that this makes intuitive sense since the teacher space entropy, according to which MTSE queries are selected, contains no feedback about the progress of the student in learning the required generalization task, and thus MTSE queries cannot be guaranteed to have a positive effect.

It is tempting to think that sufficient feedback might be restored by selecting queries orthogonal to the weight vector of a random *student* from the post-training distribution, rather than of a random *teacher* from the version space, i.e. the posterior teacher distribution. In this case, $s(\alpha)$, R and Q_N are obtained by solving (B.5) and (B.6) together with the relation $s(\alpha) = [1 - R^2/Q_N]^{1/2}$ in a self-consistent manner. The result is a power-law decay $s(\alpha) \propto \alpha^{-3/4}$ for large α, and a diverging length of the student weight vector, $Q_N \propto \alpha^{1/2}$. From equation (2.3), this leads to a similar divergence of the average generalization error, and the generalization performance achieved by such 'heuristic feedback queries' is thus even worse than for MTSE queries. Again, an intuitive explanation of this result can be found by considering the narrowing down of the input distribution along the direction of the true teacher w_V^0 that is generated by querying. For MTSE queries, this narrowing down is exponentially fast, effectively 'freezing' the length of the student weight vector to a suboptimal value for sufficiently large α, whereas for the heuristic feedback queries considered above the narrowing down is sufficiently slow to allow the length of the student weight vector to adapt steadily and thus to grow arbitrarily large as the width of the input distribution shrinks to zero.

5. Summary and discussion

We have found in our study of an unrealizable task with a linear perceptron student and a general nonlinear perceptron teacher that queries for minimum student and teacher space entropy, respectively, have very different effects on generalization performance. Minimum student space entropy (MSSE) queries essentially have the same effect as for a linear student learning a noisy linear teacher, with the effective noise level given by the sum of the actual noise level and an additional contribution due to the fact that the student cannot learn the teacher perfectly. Hence the structure of the student space is the dominant influence on the efficacy of query learning. Minimum teacher space entropy queries (MTSE) on the other hand, which we have investigated for the case of a binary perceptron teacher, perform worse than random examples, leading to a higher generalization error even for an infinite number of training examples. This result is intuitively reasonable since the teacher space entropy contains no feedback about the progress of the student in learning the required generalization task. We have also found that such feedback cannot easily be restored by more heuristic methods of query selection similar to MTSE queries.

Our results, then, are a mixture of good and bad news for query learning for minimum entropy (i.e. maximum information gain) in unrealizable tasks. The bad news is that MTSE queries, due to a lack of feedback information about student progress, are not enough to translate significant benefits in teacher space into similar improvements of student performance and may, in fact, yield worse performance than random examples. The good news is that for MSSE queries, we have found evidence that the structure of the student space is the key factor in determining the efficacy of query learning. If this result holds more generally, then statements about the benefits of query learning can be made on the basis of *how one is trying to learn* only, independently of *what one is trying to learn*—a result of obvious practical significance.

Acknowledgments

The author wishes to thank David Saad and David Barber for helpful discussions and carefully reading the manuscript.

Appendix A. Calculation for random examples and MSSE queries

In this appendix, we outline the calculation of the average generalization error for random examples and MSSE queries. For this purpose, as pointed out in section 2, it is sufficient to obtain the averages of the overlap parameters R and $Q_{\mathcal{N}}$. The averages over the Gaussian post-training distribution are straightforward and yield

$$\langle R\rangle_{P(\mathcal{N}|\Theta^{(p)})} = \frac{1}{N}\boldsymbol{w}_{\mathcal{V}}^{T}\boldsymbol{M}_{\mathcal{N}}^{-1}\boldsymbol{a} \qquad \langle Q_{\mathcal{N}}\rangle_{P(\mathcal{N}|\Theta^{(p)})} = \frac{1}{N}\boldsymbol{a}^{T}\boldsymbol{M}_{\mathcal{N}}^{-2}\boldsymbol{a} + T\frac{1}{N}\operatorname{tr}\boldsymbol{M}_{\mathcal{N}}^{-1}. \tag{A.1}$$

Since both for random examples and for MSSE queries, each new training input depends, at most, on the previous training inputs, we can use Bayes' theorem to decompose the remaining average over training sets and teachers into one over training outputs, teachers and training inputs. Formally, one has

$$P(\Theta^{(p)}|\mathcal{V})P(\mathcal{V}) = P(\{y^{\mu}\}|\{\boldsymbol{x}^{\mu}\},\mathcal{V})P(\mathcal{V})P(\{\boldsymbol{x}^{\mu}\}) \tag{A.2}$$

where

$$P(\{y^{\mu}\}|\{\boldsymbol{x}^{\mu}\},\mathcal{V}) = \prod_{\mu=1}^{p} P(y^{\mu}|\boldsymbol{x}^{\mu},\mathcal{V})$$

and we will perform the averages on the RHS of (A.2) in the order from left to right. The average over the y^{μ}-dependent terms in (A.1) yields, from the assumptions (2.1) and (2.2)

$$\langle \boldsymbol{a}\boldsymbol{w}_{\mathcal{V}}^{T}\rangle_{P(\{y^{\mu}\}|\{\boldsymbol{x}^{\mu}\},\mathcal{V})} = \frac{1}{\sqrt{N}}\sum_{\mu}\boldsymbol{x}^{\mu}\boldsymbol{w}_{\mathcal{V}}^{T}\bar{g}(h^{\mu})$$

$$\langle \boldsymbol{a}\boldsymbol{a}^{T}\rangle_{P(\{y^{\mu}\}|\{\boldsymbol{x}^{\mu}\},\mathcal{V})} = \frac{1}{N}\sum_{\mu\neq\nu}\boldsymbol{x}^{\mu}(\boldsymbol{x}^{\nu})^{T}\bar{g}(h^{\mu})\bar{g}(h^{\nu}) + \frac{1}{N}\sum_{\mu}\boldsymbol{x}^{\mu}(\boldsymbol{x}^{\mu})^{T}(\bar{g}^{2}(h^{\mu}) + \Delta^{2}(h^{\mu}))$$

where we have set $h^{\mu} = \frac{1}{\sqrt{N}} \boldsymbol{w}_{\mathcal{V}}^{T} \boldsymbol{x}^{\mu}$. Performing the average over the prior teacher distribution $P(\mathcal{V}) \propto \exp(-\frac{1}{2}\boldsymbol{w}_{\mathcal{V}}^{2}/\sigma_{\mathcal{V}}^{2})$ for *fixed* $\{\boldsymbol{x}^{\mu}\}$, the h^{μ} become Gaussian random variables with zero means and (co-)variances

$$\langle h^{\mu} h^{\nu} \rangle_{P(\mathcal{V})} = \frac{\sigma_{\mathcal{V}}^{2}}{N} (\boldsymbol{x}^{\mu})^{T} \boldsymbol{x}^{\nu} .$$

For the assumed spherical input distribution, $\frac{1}{N}(\boldsymbol{x}^{\mu})^{2} = \sigma_{x}^{2}$ and the variance of each of the h^{μ} is thus identical to $\sigma_{\mathcal{V}}^{2}\sigma_{x}^{2}$. The covariance between h^{μ} and h^{ν} for $\mu \neq \nu$ is much smaller since, for random examples, $\frac{1}{N}(\boldsymbol{x}^{\mu})^{T}\boldsymbol{x}^{\nu}$ is $\mathrm{O}(1/\sqrt{N})$. The same holds for MSSE queries, due to the pseudo-random overlaps between training inputs that they produce. The resulting weak correlation of h^{μ} and h^{ν} can be used to expand the average of $\bar{g}(h^{\mu})\bar{g}(h^{\nu})$. To this end, one writes h^{ν} as $h^{\nu} = \epsilon h^{\mu} + (1-\epsilon^{2})^{1/2}\tilde{h}$, where $\epsilon = \langle h^{\mu} h^{\nu}\rangle / \langle (h^{\mu})^{2}\rangle = (\boldsymbol{x}^{\mu})^{T}\boldsymbol{x}^{\nu}/(N\sigma_{x}^{2})$ and $\tilde{h}$ is a zero mean Gaussian variable uncorrelated with h^{μ} which has variance $\langle \tilde{h}^{2}\rangle = \langle (h^{\mu})^{2}\rangle = \langle (h^{\nu})^{2}\rangle = \sigma_{\mathcal{V}}^{2}\sigma_{x}^{2}$. Expanding in the small parameter $\epsilon = \mathrm{O}(1/\sqrt{N}) \ll 1$, one obtains ($\bar{g}' \equiv \mathrm{d}\bar{g}/\mathrm{d}h$)

$$\langle \bar{g}(h^{\mu})\bar{g}(h^{\nu})\rangle_{P(\mathcal{V})} = \langle \bar{g}(h)\rangle_{h}^{2} + \frac{1}{N\sigma_{x}^{2}} (\boldsymbol{x}^{\mu})^{T}\boldsymbol{x}^{\nu} \langle h\bar{g}(h)\rangle_{h} \langle \bar{g}'(h)\rangle_{h} + \mathrm{O}(1/N)$$

where h is a zero mean Gaussian random variable with variance $\sigma_{\mathcal{V}}^{2}\sigma_{x}^{2}$. The remaining averages over the teacher prior $P(\mathcal{V})$ are straightforward:

$$\langle \bar{g}^{2}(h^{\mu}) + \Delta^{2}(h^{\mu})\rangle_{P(\mathcal{V})} = \langle \bar{g}^{2}(h) + \Delta^{2}(h)\rangle_{h} \qquad \langle \boldsymbol{w}_{\mathcal{V}}\bar{g}(h^{\mu})\rangle_{P(\mathcal{V})} = \frac{\boldsymbol{x}^{\mu}}{\sqrt{N}\sigma_{x}^{2}} \langle h\bar{g}(h)\rangle_{h}$$

where the second equality follows from the fact that due to the isotropy of the teacher prior, the contribution from the components of $\boldsymbol{w}_{\mathcal{V}}$ orthogonal to $\boldsymbol{x}^{\mu}$ vanishes.

Collecting the results obtained so far we have for the averages of R and $Q_{\mathcal{N}}$ at fixed $\{\boldsymbol{x}^{\mu}\}$:

$$\langle R\rangle|_{\{\boldsymbol{x}^{\mu}\}} = \langle h\bar{g}(h)\rangle_{h} \frac{1}{N^{2}\sigma_{x}^{2}} \sum_{\mu} (\boldsymbol{x}^{\mu})^{T} \boldsymbol{M}_{\mathcal{N}}^{-1} \boldsymbol{x}^{\mu} \tag{A.3}$$

$$\begin{aligned} \langle Q_{\mathcal{N}}\rangle|_{\{\boldsymbol{x}^{\mu}\}} = {} & T\frac{1}{N}\operatorname{tr}\boldsymbol{M}_{\mathcal{N}}^{-1} + \frac{1}{N^{2}}\langle \bar{g}^{2}(h) + \Delta^{2}(h)\rangle_{h} \sum_{\mu} (\boldsymbol{x}^{\mu})^{T}\boldsymbol{M}_{\mathcal{N}}^{-2}\boldsymbol{x}^{\mu} \\ & + \frac{1}{N^{3}\sigma_{x}^{2}} \langle h\bar{g}(h)\rangle_{h} \langle \bar{g}'(h)\rangle_{h} \sum_{\mu\neq\nu} (\boldsymbol{x}^{\mu})^{T}\boldsymbol{M}_{\mathcal{N}}^{-2}\boldsymbol{x}^{\nu}(\boldsymbol{x}^{\mu})^{T}\boldsymbol{x}^{\nu} \\ & + \frac{1}{N^{2}}\langle \bar{g}(h)\rangle_{h}^{2} \sum_{\mu\neq\nu} (\boldsymbol{x}^{\mu})^{T}\boldsymbol{M}_{\mathcal{N}}^{-2}\boldsymbol{x}^{\nu} . \end{aligned} \tag{A.4}$$

The last term in (A.4) can be shown to vanish upon averaging over the training inputs, due to the fact that both for random examples and for MSSE queries the distribution of each individual training input $\boldsymbol{x}^{\mu}$ is invariant under the reflection $\boldsymbol{x}^{\mu} \to -\boldsymbol{x}^{\mu}$, whatever the values of the other training inputs. The summations over μ and ν in (A.3) and (A.4) can be written more succinctly by exploiting the definitions (2.8):

$$\frac{1}{N^{2}}\sum_{\mu}(\boldsymbol{x}^{\mu})^{T}\boldsymbol{M}_{\mathcal{N}}^{-k}\boldsymbol{x}^{\mu} = \frac{1}{N}\operatorname{tr}\boldsymbol{M}_{\mathcal{N}}^{-k}\boldsymbol{A} \qquad (k = 1, 2)$$

$$\begin{aligned} \frac{1}{N^{3}}\sum_{\mu\neq\nu}(\boldsymbol{x}^{\mu})^{T}\boldsymbol{M}_{\mathcal{N}}^{-2}\boldsymbol{x}^{\nu}(\boldsymbol{x}^{\mu})^{T}\boldsymbol{x}^{\nu} &= \frac{1}{N^{2}}\sum_{\mu}\operatorname{tr}\boldsymbol{M}_{\mathcal{N}}^{-2}\left[\boldsymbol{A} - \frac{1}{N}\boldsymbol{x}^{\mu}(\boldsymbol{x}^{\mu})^{T}\right]\boldsymbol{x}^{\mu}(\boldsymbol{x}^{\mu})^{T} \\ &= \frac{1}{N}\operatorname{tr}\boldsymbol{M}_{\mathcal{N}}^{-2}\boldsymbol{A}^{2} - \frac{\sigma_{x}^{2}}{N}\operatorname{tr}\boldsymbol{M}_{\mathcal{N}}^{-2}\boldsymbol{A} . \end{aligned}$$

If we now introduce the function $G(\lambda) = \frac{\sigma_x^2}{N}\langle \mathrm{tr}\, \boldsymbol{M}_{\mathcal{N}}^{-1}\rangle_{P(\{\boldsymbol{x}^\mu\})}$ and the constants γ_{eff}, σ^2_{eff} defined in (3.3) and (3.4), we can use the relations $\boldsymbol{A} = \boldsymbol{M}_{\mathcal{N}} - \lambda\sigma_x^2\mathbf{1}$ and $\partial G/\partial\lambda \equiv G' = -\frac{\sigma_x^4}{N}\langle \mathrm{tr}\, \boldsymbol{M}_{\mathcal{N}}^{-2}\rangle_{P(\{\boldsymbol{x}^\mu\})}$ to write the final averages of R and $Q_{\mathcal{N}}$ as

$$\begin{aligned}\langle R\rangle &= \sigma_v^2\gamma_{\mathrm{eff}}(1-\lambda G)\\ \langle Q_{\mathcal{N}}\rangle &= \sigma_v^2\left[\frac{T}{\sigma_x^2\sigma_v^2}G + \frac{\sigma^2_{\mathrm{eff}}}{\sigma_x^2\sigma_v^2}(G+\lambda G') + \gamma_{\mathrm{eff}}(1-2\lambda G-\lambda^2 G')\right].\end{aligned}$$

Inserting these results into (2.3), we finally obtain the expression (3.2) for the average generalization error. Parenthetically, we note that for random examples, the result (3.2) can also be obtained from a replica calculation [25].

Appendix B. Calculation for MTSE queries

In this appendix, we sketch the calculation of the average generalization error achieved by our linear perceptron student when learning to approximate a noise-free binary perceptron teacher from MTSE queries. We use the approximation explained before equation (4.1) in order to carry out the average over training inputs. Specifically, we assume that the effect of MTSE queries on the distribution of training inputs is non-negligible only for the input components along the direction of the true teacher $\boldsymbol{w}_v^0$, which are distributed according to (4.1). The other input components, i.e. the ones orthogonal to the true teacher, which for the $(p+1)$th query $\boldsymbol{x}^{p+1}$ are given by $\boldsymbol{x}_\perp^{p+1} = \boldsymbol{x}^{p+1} - x_0^{p+1}\boldsymbol{w}_v^0/|\boldsymbol{w}_v^0|$, are therefore distributed as for random examples, obeying the spherical constraint $\boldsymbol{x}^2 = N$ (remember that we set $\sigma_v^2 = \sigma_x^2 = 1$):

$$P\big(\boldsymbol{x}_\perp^{p+1}|x_0^{p+1}\big) \propto \delta\big((\boldsymbol{x}_\perp^{p+1})^2 + (x_0^{p+1})^2 - N\big).$$

In the thermodynamic limit, this spherical distribution can be replaced by a Gaussian distribution yielding the same average value of $(\boldsymbol{x}_\perp^{p+1})^2$, and the term $(x_0^{p+1})^2$, which is of order unity, can be neglected compared to N. Combining this with (4.1), the distribution of $\boldsymbol{x}^{p+1}$ can be written as a Gaussian with reduced covariance along the direction of the true teacher $\boldsymbol{w}_v^0$

$$P(\boldsymbol{x}^{p+1}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{x}^{p+1})^T\left[\mathbf{1} + (s_p^2-1)\frac{\boldsymbol{w}_v^0(\boldsymbol{w}_v^0)^T}{(\boldsymbol{w}_v^0)^2}\right]^{-1}\boldsymbol{x}^{p+1}\right\}. \tag{B.1}$$

As explained in the text, s_p, the sine of the angle between the true teacher and a random teacher from the version space defined by the first p training examples, is self-averaging in the thermodynamic limit and can therefore be regarded as a fixed constant whose value will be calculated later.

In the first part of the calculation, the average† of the teacher space entropy over all training sets generated by MTSE queries is determined, and this is then used to obtain the actual values of the s_p as explained after equation (4.2). One uses the replica trick

$$\langle \ln V\rangle_{P(\Theta^{(p)})} = \lim_{n\to 0}\frac{1}{n}\ln\langle V^n\rangle_{P(\Theta^{(p)})}$$

calculating the r.h.s. for positive integer values of n and continuing analytically to $n=0$. By introducing n replicas of the teacher space, the nth moment of the version space volume

† The teacher space entropy is, like the generalization error, self-averaging, which means that its value for a typical training set becomes arbitrarily close to its average over all training sets in the thermodynamic limit.

is expressed as

$$V^n = \int \prod_{a=1}^{n} \left(\mathrm{d}\boldsymbol{w}_{\mathrm{v}}^a \, P(\boldsymbol{w}_{\mathrm{v}}^a)\right) \prod_{\mu=1}^{p} \prod_{a=1}^{n} \Theta\left(y^\mu \frac{1}{\sqrt{N}} (\boldsymbol{w}_{\mathrm{v}}^a)^T \boldsymbol{x}^\mu\right).$$

Following [23], one can use the fact that for the noise-free binary perceptron teacher $y^\mu = \mathrm{sgn}(\frac{1}{\sqrt{N}}(\boldsymbol{w}_{\mathrm{v}}^0)^T \boldsymbol{x}^\mu)$ to decompose the product of Θ-functions for fixed training example index μ as

$$\prod_{a=1}^{n} \Theta\left(y^\mu \frac{1}{\sqrt{N}} (\boldsymbol{w}_{\mathrm{v}}^a)^T \boldsymbol{x}^\mu\right) = \prod_{a=0}^{n} \Theta\left(\frac{1}{\sqrt{N}} (\boldsymbol{w}_{\mathrm{v}}^a)^T \boldsymbol{x}^\mu\right) + \prod_{a=0}^{n} \Theta\left(-\frac{1}{\sqrt{N}} (\boldsymbol{w}_{\mathrm{v}}^a)^T \boldsymbol{x}^\mu\right).$$

Introducing Gardner representations for the Θ-functions one can rewrite this as

$$\prod_{a=0}^{n} \left(\int \frac{\mathrm{d}\hat{h}^a}{2\pi} \int_0^\infty \mathrm{d}h^a\right) \exp\left(\mathrm{i} \sum_{a=0}^{n} \hat{h}^a h^a\right) \left[\exp\left(\frac{\mathrm{i}}{\sqrt{N}} \sum_{a=0}^{n} \hat{h}^a (\boldsymbol{w}_{\mathrm{v}}^a)^T \boldsymbol{x}^\mu\right) + \mathrm{cc}\right]. \tag{B.2}$$

For a fixed true teacher $\boldsymbol{w}_{\mathrm{v}}^0$, this expression can now easily be averaged over the distribution of $\boldsymbol{x}^\mu$ as given by (B.1). In principle, an average over the distribution of true teachers, $P(\boldsymbol{w}_{\mathrm{v}}^0) \propto \exp(-\frac{1}{2}\boldsymbol{w}_{\mathrm{v}}^2)$ also has to be carried out. However, this average can be dropped due to the isotropy of the problem both in input space and in weight vector space: The result for fixed $\boldsymbol{w}_{\mathrm{v}}^0$ can only depend on $(\boldsymbol{w}_{\mathrm{v}}^0)^2$, which for the chosen Gaussian teacher space prior equals N up to corrections which can be neglected in the thermodynamic limit. Using this, the average of (B.2) over $\boldsymbol{x}^\mu$ becomes

$$2\exp\left\{-\frac{1}{2}\left[s_{\mu-1}^2 + 2s_{\mu-1}^2 \sum_{a=1}^{n} r^a + \sum_{a,b=1}^{n} \left(q^{ab} + (s_{\mu-1}^2 - 1) r^a r^b\right)\right]\right\}$$

where we have introduced the order parameters

$$r^a = \frac{1}{N} (\boldsymbol{w}_{\mathrm{v}}^a)^T \boldsymbol{w}_{\mathrm{v}}^0 \qquad q^{ab} = \frac{1}{N} (\boldsymbol{w}_{\mathrm{v}}^a)^T \boldsymbol{w}_{\mathrm{v}}^b.$$

The calculation from this point onwards proceeds exactly as in [23], yielding a saddle-point integral over r^a, q^{ab} and the corresponding conjugate order parameters. Assuming a replica-symmetric saddle-point, $r^a = r$ and $q^{ab} = q + (1-q)\delta_{ab}$, and replacing the s_p by a continuous function $s(\alpha)$ of $\alpha = p/N$, one obtains the average teacher space entropy in the form (4.2) given in the text†.

In the second part of the calculation, the average generalization error achieved by the linear perceptron student when learning from MTSE queries is calculated. The necessary averages of the overlap parameters R and Q_N can again be obtained from a replica calculation. One starts from the free energy corresponding to the Gibbs post-training distribution of students (2.7)

$$f = -\frac{T}{N} \ln Z \qquad Z = \int \mathrm{d}\boldsymbol{w}_N \exp(-E/T) \tag{B.3}$$

which can be regarded as a generating function for the averages of the overlap parameters. The free energy is self-averaging and its value in the thermodynamic limit can hence be obtained by averaging over all training sets, again using the replica method. The calculation

† Note that within an exact treatment not relying on the approximation explained before (4.1), it can be shown that the exact symmetry $q = r$ must hold at the saddle point. In our approximation, this q–r symmetry is violated. However, the violations are relatively small, in the sense that the relative deviation between q and r (and $1-q$ and $1-r$, which are the more relevant quantities for large α, when both q and r tend to unity) is never larger than 10%.

follows closely the standard method [13], with appropriate modifications taking into account the presence of a weight decay [8] and the nonlinearity of the teacher output function [15]. The only major difference from the calculation for learning from random examples is the modified input distribution (B.1). Introducing the averages

$$\overline{s}(\alpha) = \int_0^\alpha d\alpha'\, s(\alpha') \qquad \overline{s^2}(\alpha) = \int_0^\alpha d\alpha'\, s^2(\alpha') \tag{B.4}$$

one obtains the average free energy as the saddle point of

$$\frac{1}{2}\left\{\lambda Q_{\mathcal{N}} - T\frac{Q_{\mathcal{N}} - R^2}{Q_{\mathcal{N}} - Q} - T\ln[2\pi(Q_{\mathcal{N}} - Q)] + \alpha T\ln[1 + (Q_{\mathcal{N}} - Q)/T] \right.$$
$$\left. + \frac{\alpha(Q - R^2 + 1) - 2(2/\pi)^{1/2}\overline{s}(\alpha)R + \overline{s^2}(\alpha)R^2}{1 + (Q_{\mathcal{N}} - Q)/T}\right\}$$

with respect to R. $Q_{\mathcal{N}}$ and $Q = \frac{1}{N}\langle \boldsymbol{w}_{\mathcal{N}}\rangle^2_{P(\mathcal{N}|\Theta^{(p)})}$. The saddle-point values of R and $Q_{\mathcal{N}}$ are, in the thermodynamic limit, identical to their averages. Solving the saddle-point equations and restricting attention to the limit $T \to 0$, one thus finds:

$$\langle R\rangle = \left(\frac{2}{\pi}\right)^{1/2} \overline{s}(\alpha)\, \frac{F}{1+G} \tag{B.5}$$

$$\langle Q_{\mathcal{N}}\rangle = G + \lambda\frac{\partial G}{\partial\lambda} + \frac{2}{\pi}\left(\frac{\overline{s}(\alpha)}{1+G}\right)^2 \left[\frac{2F}{1+G}\frac{\partial G}{\partial\lambda} - \frac{\partial F}{\partial\lambda}\right]. \tag{B.6}$$

Here the functions G and F are given, respectively, by (3.7) and

$$\frac{1}{F} = \lambda + \frac{\overline{s^2}(\alpha)}{1+G}.$$

The average generalization error achieved by the linear student as shown in figure 5 is obtained by inserting the results (B.5) and (B.6) into (2.3) (with the substitutions $\bar{g}(h) = \mathrm{sgn}(h)$ and $\Delta^2(h) \equiv 0$ appropriate for a noise-free binary perceptron teacher) and using the numerical results for $s(\alpha)$ obtained from the calculation of the teacher space entropy. Note that (B.5) and (B.6) can also be obtained within the response function formalism of [24]. The function F then emerges as a generalization of the standard response function G in the form $F = \frac{1}{N}\langle \mathrm{tr}\, \boldsymbol{M}_s^{-1}\boldsymbol{M}_{\mathcal{N}}^{-1}\rangle_{P(\{\boldsymbol{x}^\mu\})}$. The matrix $\boldsymbol{M}_s = \lambda_s \boldsymbol{1} + \frac{1}{N}\sum_\mu (1/s_\mu^2 - 1)\boldsymbol{x}^\mu(\boldsymbol{x}^\mu)^T$, with λ_s determined by the condition $\frac{1}{N}\mathrm{tr}\, \boldsymbol{M}_s^{-1} = 1$, occurs in the correlations of the variables $z^\mu = (\boldsymbol{x}^\mu)^T\boldsymbol{w}_\mathcal{V}^0/|\boldsymbol{w}_\mathcal{V}^0|$ in the form $\langle z^\mu z^\nu\rangle_{P(\mathcal{V}|\{\boldsymbol{x}^\mu\})} = \frac{1}{N}\boldsymbol{x}^\mu \boldsymbol{M}_s^{-1}\boldsymbol{x}^\nu$.

Finally, the replica formalism can also be used to obtain the average training error achieved by MTSE queries. From the definitions (2.6) and (B.3), one has

$$\epsilon_t = \langle E_t\rangle/p = \langle E - \tfrac{1}{2}\lambda \boldsymbol{w}_{\mathcal{N}}^2\rangle/p = \frac{1}{\alpha}\left[\frac{\partial(\langle f\rangle/T)}{\partial(1/T)} - \frac{1}{2}\lambda\langle Q_{\mathcal{N}}\rangle\right].$$

By differentiating (B.5) and inserting the saddle-point value of Q, given by $Q = Q_{\mathcal{N}} - TG$, one obtains, in the limit, $T \to 0$

$$\epsilon_t = \frac{1}{2(1+G)} - \frac{\lambda}{2\alpha}(Q_{\mathcal{N}} - R^2) + \frac{\overline{s^2}(\alpha)R^2 - 2\sqrt{2/\pi}\,\overline{s}(\alpha)R}{2\alpha(1+G)}.$$

In the limit $\alpha \to \infty$, only the first term survives and converges to 1/2 since $G \to 0$; this proves the λ-independence of the asymptotic value of the average training error referred to in section 4.

References

[1] Watkin T L H, Rau A and Biehl M 1993 The statistical-mechanics of learning a rule *Rev. Mod. Phys.* **65** 499–556

[2] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City, CA: Addison-Wesley)

[3] Seung H S, Opper M and Sompolinsky H 1992 Query by committee *Proc. Fifth Annual ACM Workshop on Computational Learning Theory (COLT '92) (Pittsburgh)* (New York: ACM) pp 287–94

[4] Freund Y, Seung H S, Shamir E and Tishby N 1993 Information, prediction, and query by committee *Advances in Neural Information Processing Systems* vol 5 ed S J Hanson, J D Cowan and C Lee Giles (San Mateo, CA: Kaufmann) pp 483–90

[5] Sollich P 1994 Query construction, entropy, and generalization in neural network models *Phys. Rev.* E **49** 4637–51

[6] Sollich P and Saad D 1995 Learning from queries for maximum information gain in imperfectly learnable problems *Advances in Neural Information Processing Systems* vol 7 ed G Tesauro, D S Touretzky and T K Leen (Cambridge, MA: MIT Press) pp 287–94

[7] Krogh A and Hertz J A 1992 Generalization in a linear perceptron in the presence of noise *J. Phys. A: Math. Gen.* **25** 1135–47

[8] Dunmur A P and Wallace D J 1993 Learning and generalization in a linear perceptron stochastically trained with noisy data *J. Phys. A: Math. Gen.* **26** 5767–79

[9] Barber D, Saad D and Sollich P 1995 Finite-size effects and optimal test set size in linear perceptrons *J. Phys. A: Math. Gen.* **28** 1325–34

[10] Levin E, Tishby N and Solla S A 1990 A statistical approach to learning and generalization in layered neural networks *Proc. IEEE* **78** 1568–74

[11] Bruce A and Saad D 1994 Statistical mechanics of hypothesis evaluation *J. Phys. A: Math. Gen.* **27** 3355–63

[12] Baldi P F and Hornik K 1995 Learning in linear neural networks—a survey *IEEE Trans. Neural Net.* **6** 837–858

[13] Seung H S, Sompolinsky H and Tishby N 1992 Statistical-mechanics of learning from examples *Phys. Rev.* A **45** 6056–91

[14] Krogh A and Hertz J A 1992 A simple weight decay can improve generalization *Advances in Neural Information Processing Systems* vol 4 ed J E Moody, S J Hanson and R P Lippmann (San Mateo, CA: Kaufmann) pp 950–7

[15] Bös S, Kinzel W and Opper M 1993 Generalization ability of perceptrons with continuous outputs *Phys. Rev.* E **47** 1384–91

[16] Holmström L and Koistinen P 1992 Using additive noise in back-propagation training *IEEE Trans. Neural Net.* **3** 24–38

[17] Gardner E J, Stroud N and Wallace D J 1989 Training with noise and the storage of correlated patterns in a neural network model *J. Phys. A: Math. Gen.* **22** 2019–30

[18] Wong K Y M and Sherrington D 1993 Neural networks optimally trained with noisy data *Phys. Rev.* E **47** 4465–82

[19] Pilz J 1991 *Bayesian Estimation and Experimental Design in Linear Regression Models* 2nd edn (Chichester: Wiley) (1983 1st edn (Leipzig: Teubner))

[20] MacKay D J C 1992 Information-based objective functions for active data selection *Neural Comput.* **4** 590–604

[21] Silvey S D 1980 *Optimal design* (London: Chapman and Hall)

[22] Watkin T L H and Rau A 1992 Selecting examples for perceptrons *J. Phys. A: Math. Gen.* **25** 113–21

[23] Györgyi G and Tishby N 1990 Statistical theory of learning a rule *Neural Networks and Spin Glasses* ed W Theumann and R Köberle (Singapore: World Scientific) pp 3–36

[24] Sollich P 1994 Finite-size effects in learning and generalization in linear perceptrons *J. Phys. A: Math. Gen.* **27** 7771–84

[25] Dunmur A P 1995 Nonlinear rule learning by a simple perceptron, unpublished