On-Line Learning with Restricted Training Sets: An Exactly Solvable Case

H.C. Rae, P. Sollich and A.C.C. Coolen

Department of Mathematics
King's College, University of London
Strand, London WC2R 2LS, U.K.

Abstract

We solve the dynamics of on-line Hebbian learning in large perceptrons exactly, for the regime where the size of the training set scales linearly with the number of inputs. We consider both noiseless and noisy teachers. Our calculation cannot be extended to non-Hebbian rules, but the solution provides a convenient and welcome benchmark with which to test more general and advanced theories for solving the dynamics of learning with restricted training sets.

PACS: 87.10.+e, 02.50.-r. To appear in *Journal of Physics A*.

1 Introduction

Considerable progress has been made in understanding the dynamics of online learning in layered artificial neural networks through the application of the methods of statistical mechanics; see e.g. [1, 2, 3]. For an excellent review of the state-of-the-art in the field, we refer the reader to the workshop proceedings [4], where a large number of references to earlier work can also be found; the special case of binary perceptrons has been surveyed in detail in [5]. For the most part, theoretical work has concentrated on systems where the training set is much larger than the number of weight updates. In such circumstances the probability that any given question will be repeated during the training process is negligible and it is possible to assume for large networks, via the central limit theorem, that their local field distribution is always Gaussian. In this paper we consider *restricted training sets*; we suppose that the size p of the training set scales linearly with N, the number of inputs. As a consequence the probability that a question will reappear during the training process is no longer negligible, the assumption that the local fields have Gaussian distributions is not tenable, and it is clear that correlations will develop between the weights and the questions in the training set as training progresses. In fact, the non-Gaussian character of the local fields should be a *prediction* of any satisfactory theory of learning with restricted training sets, as this is clearly demanded by numerical simulations.

Several authors [6, 7, 8, 9, 10, 11] have discussed learning with restricted training sets but constructing a general theory is difficult. A simple model of learning with restricted training sets which can be solved *exactly* is therefore particularly attractive and provides a yardstick against which more difficult and sophisticated general theories can, in due course, be tested and compared. We show how this can be accomplished for on-line Hebbian learning in perceptrons with restricted training sets and we obtain exact solutions for the generalization error, the training error and the field distribution for a class of noisy teacher networks and student networks with arbitrary weight decay. We work out in detail the two particular but representative cases of output noise and Gaussian weight noise. Our theory is found to be in excellent agreement with numerical simulations and our predictions for the probability density of the student field are a striking confirmation of them, making it clear that we are indeed dealing with local fields which are non-Gaussian. An outline of our results is to appear in the conference proceedings [12].

2 Definitions and Explicit Microscopic Expressions

We study on-line learning in a student perceptron S, which tries to learn a task defined by a noisy teacher perceptron T. The student input-output mapping is specified by a weight vector $\boldsymbol{J}$ according to

$$S : \{-1,1\}^N \to \{-1,1\} \qquad S(\boldsymbol{\xi}) = \mathrm{sgn}[\boldsymbol{J}\cdot\boldsymbol{\xi}]\ .$$

For a given $\boldsymbol{J}$, this is a deterministic mapping from binary inputs to binary outputs. The teacher output $T(\boldsymbol{\xi})$, on the other hand, is stochastic. In its most general form, it is determined by the probabilities $P(T=\pm 1|\boldsymbol{\xi})$. These are related to the *average* teacher output $\overline{T}(\boldsymbol{\xi})$ for a given input $\boldsymbol{\xi}$ by

$$P(T=\pm 1|\boldsymbol{\xi}) = \frac{1}{2}[1\pm\overline{T}(\boldsymbol{\xi})], \qquad \text{or} \qquad P(T|\boldsymbol{\xi}) = \frac{1}{2}[1+T\overline{T}(\boldsymbol{\xi})]\ . \tag{1}$$

To ensure that this noisy teacher mapping can be thought of as the corrupted output of an underlying 'clean' perceptron with weights $\boldsymbol{B}^*$, we make the mild assumption that the average teacher output can be written in the form

$$\overline{T}(\boldsymbol{\xi}) = \tau(y), \qquad y = \boldsymbol{B}^*\cdot\boldsymbol{\xi} \tag{2}$$

with some function $\tau(y)$. In other words, the noise process preserves, on average, the perceptron structure of the teacher. The uncorrupted teacher weight vector is taken to be normalized such that $(\boldsymbol{B}^*)^2 = 1$, with each component B_i^* of $\mathcal{O}(N^{-\frac{1}{2}})$. We also assume that inputs are sampled randomly from a uniform distribution[1] on $\{-1,1\}^N$. Typical values of the (uncorrupted) 'teacher field' y are then of $\mathcal{O}(1)$; in the thermodynamic limit $N\to\infty$ that we will be interested in, y is Gaussian with zero mean and unit variance.

The class of noise processes allowed by (2) is quite large and includes the standard cases of output noise and Gaussian weight noise that are often discussed in the literature. For output noise, the sign of the clean teacher output $\mathrm{sgn}(y)$ is inverted with probability λ, i.e.,

$$P(T|\boldsymbol{\xi}) = (1-\lambda)\ \theta(Ty) + \lambda\theta(-Ty), \qquad \tau(y) = (1-2\lambda)\mathrm{sgn}(y) \tag{3}$$

For Gaussian weight noise, the teacher output is produced from a corrupted teacher weight vector $\boldsymbol{B}$. The corrupted weights $\boldsymbol{B}$ differ from $\boldsymbol{B}^*$ by the addition of Gaussian noise of standard

[1] This choice of input distribution is not critical. In fact, any other distribution with $\langle\xi_i\rangle = 0$ and $\langle\xi_i\xi_j\rangle = \delta_{ij}$ will give results identical to the ones for the present case in the limit $N\to\infty$. Examples would be real-valued inputs with either a Gaussian distribution with zero mean and unit variance for each component, or a uniform distribution over the hypersphere $\boldsymbol{\xi}^2 = N$. Likewise, we only actually require that assumption (2) should hold with probability one (i.e., for almost all inputs) in the limit $N\to\infty$.

deviation $\Sigma/\sqrt{N}$ to each component, i.e.,

$$P(\boldsymbol{B}) = \left[\frac{N}{2\pi\Sigma^2}\right]^{N/2} \exp\left(-\frac{N}{2\Sigma^2}(\boldsymbol{B}-\boldsymbol{B}^*)^2\right). \tag{4}$$

The scaling with N here is chosen to get a sensible result in the thermodynamic limit (corrupted and clean weights clearly need to be of the same order). The corrupted teacher field is then $z = \boldsymbol{B}\cdot\boldsymbol{\xi} = y + \Delta$, with Δ a Gaussian random variable with zero mean and variance Σ^2, and hence

$$\tau(y) = \langle \mathrm{sgn}(y+\Delta)\rangle_{\Delta} = \mathrm{erf}(y/\sqrt{2}\Sigma). \tag{5}$$

In the numerical examples presented later, we will focus on the above two noise models. But our analytical treatment applies to any teacher that is compatible with the assumption (2). This covers, for example, the more complex cases of 'reversed wedge' teachers (where $\tau(y) = \mathrm{sgn}(y)$ for $|y| > d$ and $\tau(y) = -\mathrm{sgn}(y)$ otherwise, d being the wedge 'thickness') and noisy generalizations of these.

Our learning rule will be the on-line Hebbian rule, i.e.

$$\boldsymbol{J}(\ell+1) = \left(1-\frac{\gamma}{N}\right)\boldsymbol{J}(\ell) + \frac{\eta}{N}\,\boldsymbol{\xi}^{\mu(\ell)}T^{\mu(\ell)} \tag{6}$$

where the non-negative parameters γ and η are the weight decay and the learning rate, respectively. Learning starts from an initial set of student weights $\boldsymbol{J}_0 \equiv \boldsymbol{J}(0)$, for which we assume (as for the teacher weights) that $J_i(0) = \mathcal{O}(N^{-\frac{1}{2}})$. At each iteration step ℓ a training example, comprising an input vector $\boldsymbol{\xi}^{\mu(\ell)}$ and the corresponding teacher output $T^{\mu(\ell)}$, is picked at random (with replacement) from the *training set* D. This training set consists of $p = \alpha N$ examples, $D = \{(\boldsymbol{\xi}^{\mu}, T^{\mu}),\ \mu = 1\ldots p\}$; it remains unchanged throughout the learning process. Each training input vector $\boldsymbol{\xi}^{\mu}$ is assumed to be randomly drawn from $\{-1,1\}^N$ (independently of other training inputs, and of $\boldsymbol{J}_0$ and $\boldsymbol{B}^*$), and the output $T^{\mu} = T(\boldsymbol{\xi}^{\mu})$ provided by the noisy teacher. We call this kind of scenario 'consistent noise': To each training input corresponds a single output value which is produced by the teacher once and for all before learning begins; the teacher is *not* asked to produce new noisy outputs each time a training input is selected for a weight update.

There are two sources of randomness in the above scenario. First of all there is the random realization of the 'path' $(\mu(0), \mu(1), \mu(\ell), \ldots)$. This is simply the dynamic randomness of the stochastic process that gives the evolution of the vector $\boldsymbol{J}$; it arises from the random selection of examples from the training set. Averages over this process will be denoted as $\langle\ldots\rangle$. Secondly there is the randomness in the composition of the training set. We will write averages over all training sets as $\langle\ldots\rangle_{\mathrm{sets}}$. We note that

$$\left\langle f(\boldsymbol{\xi}^{\mu(\ell)}, T^{\mu(\ell)})\right\rangle = \frac{1}{p}\sum_{\mu=1}^{p} f(\boldsymbol{\xi}^{\mu}, T^{\mu}) \qquad \text{(for all } \ell)$$

and that averages over all possible realizations of the training set are given by

$$\left\langle f[(\boldsymbol{\xi}^1, \boldsymbol{B}^1), (\boldsymbol{\xi}^2, \boldsymbol{B}^2), \ldots, (\boldsymbol{\xi}^p, \boldsymbol{B}^p)]\right\rangle_{\mathrm{sets}} =$$

$$= \sum_{\boldsymbol{\xi}^1}\sum_{\boldsymbol{\xi}^2}\cdots\sum_{\boldsymbol{\xi}^p}\left(\frac{1}{2^N}\right)^p \sum_{T^1,\ldots,T^p=\pm1}\left[\prod_{\mu=1}^{p} P(T^{\mu}|\boldsymbol{\xi}^{\mu})\right] f[(\boldsymbol{\xi}^1, \boldsymbol{B}^1), (\boldsymbol{\xi}^2, \boldsymbol{B}^2), \ldots, (\boldsymbol{\xi}^p, \boldsymbol{B}^p)]$$

where $\boldsymbol{\xi}^\mu \in \{-1,1\}^N$.

Our aim is to evaluate the performance of the on-line Hebbian learning rule (6) as a function of the number of training steps m. This calculation becomes tractable in the thermodynamic limit $N \to \infty$; the appropriate time variable is then $t = m/N$. Basic quantities of interest are the generalization error and the training error. The generalization error, which we choose to measure with respect to the *clean* teacher, is the probability of student and (clean) teacher producing different outputs on a randomly chosen test input. Hence $E_\mathrm{g} = \langle \theta[-(\boldsymbol{J}\cdot\boldsymbol{\xi})(\boldsymbol{B}^*\cdot\boldsymbol{\xi})]\rangle_{\boldsymbol{\xi}}$, with the usual result

$$E_\mathrm{g} = \frac{1}{\pi} \arccos\left(\frac{R}{\sqrt{Q}}\right). \tag{7}$$

Here $Q = \boldsymbol{J}^2$ is the squared length of the student weight vector, and $R = \boldsymbol{B}^*\cdot\boldsymbol{J}$ its overlap with the teacher weights. These are our basic scalar observables. The training error E_t is the fraction of errors that the students makes on the training set, i.e., the fraction of training outputs that are predicted incorrectly. It is given by

$$E_\mathrm{t} = \int dx \sum_{T=\pm 1} P(x,T)\,\theta(-Tx)$$

where $P(x,T)$ is the joint distribution of the student fields $x = \boldsymbol{J}\cdot\boldsymbol{\xi}$ and the teacher outputs T over the training set. Because the teacher outputs depend on the teacher fields, according to $P(T|y) = \frac{1}{2}[1 + T\tau(y)]$, it is useful to include the latter and to calculate the distribution $P(x,y,T)$; we will see later that this also leads to a rather transparent form of the result. Formally, the joint field/output distribution is defined in the obvious way,

$$P(x,y,T) = \frac{1}{p}\sum_{\mu=1}^{p} \;\delta(x - \boldsymbol{J}\cdot\boldsymbol{\xi}^\mu)\;\delta(y - \boldsymbol{B}^*\cdot\boldsymbol{\xi}^\mu)\;\delta_{T,T^\mu}\;. \tag{8}$$

For infinitely large systems, $N \to \infty$, one can prove that the fluctuations in mean-field observables such as $\{Q, R, P(x,y,T)\}$, due to the randomness in the dynamics, will vanish [10]. Furthermore one assumes, with convincing support from numerical simulations, that for $N \to \infty$ the evolution of such observables, when observed for different random realizations of the training set, will be reproducible (i.e., the sample-to-sample fluctuations will also vanish, which is called 'self-averaging'). Both properties are central ingredients of all current theories. We are thus led to the introduction of averages of our observables, both with respect to the dynamical randomness and with respect to the randomness in the training set (always to be carried out in precisely this order):

$$Q(t) = \lim_{N\to\infty} \langle\,\langle Q\rangle\,\rangle_\mathrm{sets} \qquad R(t) = \lim_{N\to\infty} \langle\,\langle R\rangle\,\rangle_\mathrm{sets} \tag{9}$$

$$P_t(x,y,T) = \lim_{N\to\infty} \langle\,\langle P(x,y,T)\rangle\,\rangle_\mathrm{sets} \tag{10}$$

The large N-limits here are taken at constant t and α, i.e., with the number of weight updates and the number of training examples scaling as $m = Nt$ and $p = N\alpha$, respectively.

Iterating the learning rule (6), we find an explicit expression for the student weight vector after m training steps:

$$\boldsymbol{J}(m) = \sigma^m \boldsymbol{J}_0 \;+\; \frac{\eta}{N}\sum_{\ell=0}^{m-1} \sigma^{m-\ell-1} \boldsymbol{\xi}^{\mu(\ell)}\, T^{\mu(\ell)} \tag{11}$$

where

$$\sigma = 1 - \frac{\gamma}{N} \, .$$

Eq. (11) will be the natural starting point for our calculation. We will also frequently encounter averages of the form

$$\langle \boldsymbol{v} \cdot \boldsymbol{\xi} \, T(\boldsymbol{\xi}) \rangle_{\boldsymbol{\xi},T}$$

which we now calculate. The average over T is trivial and, using assumption (2), gives $\langle \boldsymbol{v} \cdot \boldsymbol{\xi} \, \tau(\boldsymbol{B}^* \cdot \boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}}$. Provided all components of the vector $\boldsymbol{v}$ are of the same order, $v = \boldsymbol{v} \cdot \boldsymbol{\xi}$ and $y = \boldsymbol{B}^* \cdot \boldsymbol{\xi}$ become zero mean Gaussian variables for $N \to \infty$ with $\langle vy \rangle = \boldsymbol{v} \cdot \boldsymbol{B}^*$ and $\langle y^2 \rangle = (\boldsymbol{B}^*)^2 = 1$. By averaging over v first for fixed y, we obtain the desired result

$$\langle \boldsymbol{v} \cdot \boldsymbol{\xi} \, T(\boldsymbol{\xi}) \rangle_{\boldsymbol{\xi},T} = \rho \, \boldsymbol{v} \cdot \boldsymbol{B}^*, \qquad \rho = \langle y \tau(y) \rangle = \int Dy \, y \, \tau(y) \tag{12}$$

with the familiar short-hand $Dy = (2\pi)^{-\frac{1}{2}} e^{-y^2/2} dy$. Using (3) and (5), one finds for the proportionality constant ρ the explicit expressions

$$\rho = \sqrt{\frac{2}{\pi}} \, (1 - 2\lambda) \qquad \text{(output noise)} \tag{13}$$

$$\rho = \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{1 + \Sigma^2}} \qquad \text{(Gaussian weight noise)} \tag{14}$$

for the two noise models that we will consider in some detail.

3 Simple Scalar Observables

It is a simple matter to calculate the values of Q and R after m learning steps, using (11). For Q, we find

$$Q = \sigma^{2m} \boldsymbol{J}_0^2 + \frac{2\eta}{N} \sum_{\ell=0}^{m-1} \sigma^{2m-\ell-1} \boldsymbol{J}_0 \cdot \boldsymbol{\xi}^{\mu(\ell)} \, T^{\mu(\ell)} + \frac{\eta^2}{N^2} \sum_{\ell,\ell'=0}^{m-1} \sigma^{m-\ell-1} \sigma^{m-\ell'-1} \boldsymbol{\xi}^{\mu(\ell)} \cdot \boldsymbol{\xi}^{\mu(\ell')} \, T^{\mu(\ell)} T^{\mu(\ell')}$$

We now average both with respect to dynamical (or path) randomness and with respect to the randomness in the training set, and take the limit $N \to \infty$ at constant learning time $t = m/N$ (see (9)). Separating out the terms with $\ell = \ell'$ from the double sum, and using (12), we obtain

$$\begin{aligned} Q(t) &= e^{-2\gamma t} Q_0 + 2\eta\rho R_0 \lim_{N\to\infty} \frac{1}{N} \sum_{\ell=0}^{tN} \sigma^{2tN-\ell} + \eta^2 \lim_{N\to\infty} \frac{1}{N} \sum_{\ell=0}^{tN} \sigma^{2tN-2\ell} \\ &\quad + \eta^2 \lim_{N\to\infty} \frac{1}{N^2} \sum_{\ell \neq \ell'} \sigma^{tN-\ell} \sigma^{tN-\ell'} \left\langle \left\langle \boldsymbol{\xi}^{\mu(\ell)} \cdot \boldsymbol{\xi}^{\mu(\ell')} \, T^{\mu(\ell)} T^{\mu(\ell')} \right\rangle \right\rangle_{\text{sets}} . \end{aligned}$$

Here $Q_0 = \boldsymbol{J}_0^2$ and $R_0 = \boldsymbol{J}_0 \cdot \boldsymbol{B}^*$ are the squared length and overlap of the initial student weights, respectively. After averaging over the dynamical randomness, the average in the last term becomes $(1/p^2) \sum_{\mu,\nu=1}^{p} \langle \boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\nu \, T^\mu T^\nu \rangle_{\text{sets}}$. The terms with $\mu = \nu$ each contribute $(\boldsymbol{\xi}^\mu)^2 = N$

to this sum; the others make a contribution of ρ^2 each, as one finds by applying (12) twice. Assembling everything, we have

$$Q(t) = e^{-2\gamma t}Q_0 + 2\rho R_0 \frac{\eta}{\gamma} e^{-\gamma t}(1-e^{-\gamma t}) + \frac{\eta^2}{2\gamma}(1-e^{-2\gamma t}) + \frac{\eta^2}{\gamma^2}\left(\frac{1}{\alpha}+\rho^2\right)(1-e^{-\gamma t})^2 \qquad (15)$$

where ρ is given by equations (13, 14) in the examples of output noise and Gaussian weight noise, respectively, and more generally by (12). In a similar manner we find that

$$\begin{aligned} R(t) &= \lim_{N\to\infty} \sigma^{tN} R_0 + \frac{\eta}{N}\sum_{\ell=0}^{tN} \sigma^{tN-\ell} \left\langle\!\left\langle\, \boldsymbol{B}^* \cdot \boldsymbol{\xi}^{\mu(\ell)}\, T^{\mu(\ell)} \,\right\rangle\!\right\rangle_{\text{sets}} \\ &= e^{-\gamma t} R_0 + \frac{\eta\rho}{\gamma}(1-e^{-\gamma t}) \qquad (16) \end{aligned}$$

We note in passing that our calculations easily generalize to the case of a variable learning rate $\eta(t)$. Sums such as $\frac{\eta}{N}\sum_{\ell=0}^{tN}\sigma^{tN-\ell}$ would simply be replaced by $\frac{1}{N}\sum_{\ell=0}^{tN}\sigma^{tN-\ell}\eta(\ell/N)$. Using $\sigma^{tN-\ell} = (1-\gamma/N)^{tN-\ell} = \exp[-\gamma t + \gamma\ell/N + \mathcal{O}(1/N)]$ we see that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{\ell=0}^{tN}\sigma^{tN-\ell}\eta(\ell/N) = \int_0^t ds\; e^{-\gamma(t-s)}\eta(s)$$

which reduces to the familiar result in the case when η is constant. Other sums involving a variable learning rate can be treated in similar fashion.

The generalization error follows directly from the above results and (7); its asymptotic value is

$$E_{\mathrm{g},\infty} = \lim_{t\to\infty} E_{\mathrm{g}}(t) = \frac{1}{\pi}\arccos\left(\frac{\rho}{\sqrt{\gamma/2 + 1/\alpha + \rho^2}}\right) \qquad (17)$$

In fact, one sees from (15,16) that (for $N\to\infty$) all noisy teachers with the same ρ will give the same generalization error at any time t. This is true, in particular, of output noise and Gaussian weight noise when their respective parameters λ and Σ are related by $1-2\lambda = (1+\Sigma^2)^{-\frac{1}{2}}$. More generally, one can use (13) to associate, with any type of teacher noise obeying our basic assumption (2), an effective output noise parameter λ_{eff} given by

$$1 - 2\lambda_{\text{eff}} = \sqrt{\frac{\pi}{2}}\,\rho = \sqrt{\frac{\pi}{2}}\,\langle y\tau(y)\rangle \qquad (18)$$

Note, however, that this effective teacher error probability λ_{eff} will in general not be identical to the *real* teacher error probability λ_{real}. The latter is defined as the probability of an incorrect teacher output for a random input, $\lambda_{\text{real}} = \langle P(T = -\text{sgn}(\boldsymbol{B}^*\cdot\boldsymbol{\xi})|\boldsymbol{\xi})\rangle_{\boldsymbol{\xi}}$. Using (1), this can be rewritten as $\lambda_{\text{real}} = \left\langle \frac{1}{2}[1-\text{sgn}(\boldsymbol{B}^*\cdot\boldsymbol{\xi})\overline{T}(\boldsymbol{\xi})]\right\rangle_{\boldsymbol{\xi}}$, and with (2) one obtains

$$1 - 2\lambda_{\text{real}} = \langle \text{sgn}(y)\tau(y)\rangle\,. \qquad (19)$$

Comparing with (18), one sees that in the effective error probability that is relevant to our Hebbian learning process, errors for inputs with large teacher fields y are weighted more heavily than in the real error probability. For output noise, this is irrelevant because the probability of an incorrect teacher error is independent of y, and λ_{real} and λ_{eff} are therefore identical. For

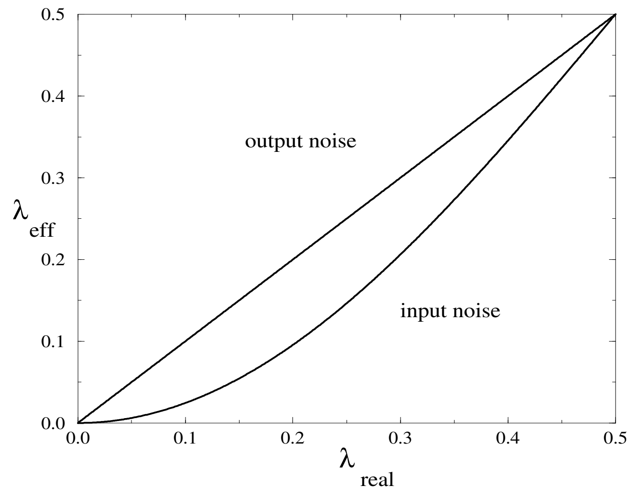

Figure 1: Relation between the real output error probability $\lambda_{\rm real}$ and the effective output error probability $\lambda_{\rm eff}$ for noisy teachers. For output noise the two are identical. One observes that, for the same 'real' noise level, weight noise is significantly less disruptive to the learning process than output noise.

Gaussian weight noise, on the other hand, errors are most likely to occur near the decision boundary of the teacher ($y = 0$). These are suppressed by the weighting in the effective error probability, and so $\lambda_{\rm eff} < \lambda_{\rm real}$. Explicitly, one finds in this case $\lambda_{\rm real} = \frac{1}{\pi}\arctan\Sigma$, and from (14), $\lambda_{\rm eff} = \frac{1}{2}[1-(1+\Sigma^2)^{-1/2}]$; the relation between effective and real error probabilities for Gaussian weight noise (see figure 1) is therefore

$$\lambda_{\rm eff} = \frac{1}{2}[1-\cos(\pi\lambda_{\rm real})] = \sin^2(\pi\lambda_{\rm real}/2).$$

Having established the general features of the evolution of the generalization error in our system, we briefly analyse some special limits in order to clarify the roles of the various parameters involved. We note first that eqs. (15,16) can be combined to give the squared length of the component of the student weight vector orthogonal to the teacher $\boldsymbol{B}^*$ as

$$Q(t) - R^2(t) = e^{-2\gamma t}(Q_0 - R_0^2) + \frac{\eta^2}{2\gamma}(1-e^{-2\gamma t}) + \frac{\eta^2}{\alpha\gamma^2}(1-e^{-\gamma t})^2. \tag{20}$$

This is independent of ρ and hence of the noise level of the teacher: Because of the perceptron structure of the teacher, the components of the training inputs orthogonal to $\boldsymbol{B}^*$ are uncorrelated with the training outputs; their influence on the learning process is therefore not modified by noise. In the short time regime, eqs. (16,20) simplify to

$$\begin{aligned} R(t) &= R_0 + \eta\rho t \\ Q(t) - R^2(t) &= Q_0 - R_0^2 + \eta^2 t + \frac{\eta^2}{\alpha}t^2. \end{aligned} \tag{21}$$

Note that the second contribution to $Q - R^2$ is of a 'diffusive' nature ($Q - R^2 \sim t$). It reflects the stochastic nature of the online learning process; correspondingly, it vanishes in the small learning rate limit ($\eta \to 0$ at constant rescaled learning time ηt) where such (path) fluctuation effects average out. The finite-α contribution, on the other hand, does survive in this limit; its t^2 scaling corresponds to 'ballistic' motion, i.e., a constant drift velocity of the student weight vector. This of course arises from the fact that the training set average of the vectors $\boldsymbol{\xi}^\mu T^\mu$ has a nonzero component orthogonal to $\boldsymbol{B}^*$ as long as α is finite.

The expressions (21) are valid in the short-time region characterized by $t \ll 1/\gamma$; they are independent of γ because in this regime the number of learning steps is still too small for the weight decay to have had a noticeable effect. For larger learning times, on the other hand, Q, R and hence also $E_{\rm g}$ approach their asymptotic values exponentially quickly with decay rate γ. For $\gamma \to 0$, this rate vanishes; eqs. (21) are then valid at all times, and the generalization error decays as a power law, $E_{\rm g}(t) - E_{{\rm g},\infty} \sim 1/t$ (or $E_{\rm g}(t) \sim 1/\sqrt{t}$ in the special case $E_{{\rm g},\infty} = 0$).

In addition to $1/\gamma$, the time scale for crossover into the asymptotic regime, there is a second time scale in the problem: When $t \approx \alpha$, the finite size of the training set will make itself felt because examples start to be 'recycled' in the learning process. Consistent with this intuition, one can show in eqs. (20,21) that the value of α does not affect the results as long as $t \ll \alpha$. If $\alpha \gg 1/\gamma$, then the system asymptotes before it even 'realizes' that α was finite, and one finds indeed that in this case the results are independent of α for all learning times t.

At this point, a brief discussion of the effect of the weight decay γ on the asymptotic generalization error $E_{{\rm g},\infty}$, eq. (17), is in order. If we think of online learning as approximating offline gradient descent on some fixed cost function, scaled by a learning rate η, then it may seem surprising that $E_{{\rm g},\infty}$ is independent of this learning rate, while the value of the weight decay γ remains relevant[2] even for $\alpha \to \infty$. However, the analogy with offline gradient descent requires a different parameterization of the weight decay, which resolves these issues. The cost function for offline (batch) gradient descent normally consists of the batch training error plus a fixed quadratic (weight decay) penalty term. If we call $\tilde{\gamma}$ the coefficient of this penalty term, then this corresponds to a value of the weight decay $\gamma = \eta\tilde{\gamma}/\alpha$ in our scenario. Here the factor η comes from the overall scaling of the update steps with η; the factor $1/\alpha$ arises because the batch training error scales with the training set size while the penalty term remains fixed (compare the discussion in Ref. [8]). With this scaling, the effect of the 'offline weight decay' $\tilde{\gamma}$ on $E_{{\rm g},\infty}$ does in fact vanish for $\alpha \to \infty$; furthermore, $E_{{\rm g},\infty}$ decreases as the learning rate η is decreased at constant $\tilde{\gamma}$ and is minimal at $\eta = 0$.

We close this section by asking whether the generalization error $E_{\rm g}(t)$ can have a minimum at a finite time t, i.e., whether overtraining can occur in our problem. After a straightforward but tedious calculation we find that $E_{\rm g}(t)$ as given by (7,15,16) is stationary at the point $t = t^*$, where

$$t^* = \frac{1}{\gamma} \log \left[\frac{-2\gamma\eta^{-1}\rho Q_0 \sin^2(\pi E_{{\rm g},0}) - 2Q_0^{1/2}\alpha^{-1}\cos(\pi E_{{\rm g},0}) + \eta\rho}{\eta\rho - Q_0^{1/2}(2/\alpha + \gamma)\cos(\pi E_{{\rm g},0})} \right] \tag{22}$$

Here $E_{{\rm g},0} \equiv E_{\rm g}(0)$ is the initial generalization performance of the student. It turns out that $E_{\rm g}(t)$ has a *minimum* at t^* if the numerator of the logarithm in equation (22) is negative. Of course, t^* must be real and *positive* – which demands that the denominator of the logarithmic term in (22) be negative, and that the numerator be less than the denominator. This implies that $E_{\rm g}(t)$ will have a minimum at t^* if

$$\eta < \rho^{-1}(\gamma + 2/\alpha)\, Q_0^{1/2} \cos(\pi E_{{\rm g},0}) \qquad \text{and} \qquad \eta < 2\rho\, Q_0^{1/2} \sin(\pi E_{{\rm g},0}) \tan(\pi E_{{\rm g},0}) \tag{23}$$

[2] To avoid confusion, we stress that while the *asymptotic* generalization error $E_{{\rm g},\infty}$ is always optimized by the choice $\gamma = 0$, this is not true in general for $E_{\rm g}(t)$ at finite learning times t.

These conditions apply when $E_{g,0} \in [0, \frac{1}{2})$, corresponding to an initial performance *better* than random guessing; when $E_{g,0} \geq \frac{1}{2}$, the generalization error $E_g(t)$ is always monotonically decreasing in time. When the conditions (23) are satisfied (which is always the case for sufficiently small learning rate η, as long as $E_{g,0} < \frac{1}{2}$) the generalization error has a minimum at t^*. Maxima in $E_g(t)$ are also possible; the corresponding conditions are obtained from (23) by reversing both inequality signs. We reiterate that the above effects can occur only if the initial performance of the student is better than random guessing. They arise essentially out of the competition between 'forgetting' the initial state, and learning a new weight vector from the training data (whose performance, given for $t \to \infty$ by $E_{g,\infty}$, may be better or worse than the initial one). In this sense, they are rather more trivial than more conventional overtraining effects observed in other systems, which occur only for noisy training sets of finite size; in our case, a non-monotonic $E_g(t)$ is possible even for noise-free teachers and infinite training sets.

4 Joint Field Distribution

The calculation of the average of the joint field distribution starting from equation (10) is more difficult than that of the scalar observables. It is convenient to work in terms of the characteristic function

$$\hat{P}_t(\hat{x}, \hat{y}, \hat{T}) = \left\langle e^{-i(\hat{x}x + \hat{y}y + \hat{T}T)} \right\rangle_{P_t(x,y,T)} ; \tag{24}$$

using equations (8,10,11), we then find that

$$\begin{aligned} \hat{P}_t(\hat{x}, \hat{y}, \hat{T}) &= \lim_{N\to\infty} \left\langle \frac{1}{p} \sum_{\mu=1}^{p} \exp[-i(\hat{x}\sigma^{tN} \boldsymbol{J}_0 \cdot \boldsymbol{\xi}^\mu + \hat{y}\boldsymbol{B}^* \cdot \boldsymbol{\xi}^\mu + \hat{T}T^\mu)] \right. \\ & \left. \times \left\langle \exp\left(-\frac{i\eta\hat{x}}{N} \sum_{\ell=0}^{tN} \sigma^{tN-\ell}\, \boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^{\mu(\ell)}\, T^{\mu(\ell)} \right) \right\rangle \right\rangle_{\text{sets}} \end{aligned} \tag{25}$$

Performing the path average gives

$$\left\langle \exp\left(-\frac{i\eta\hat{x}}{N} \sum_{\ell=0}^{tN} \sigma^{tN-\ell} \boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^{\mu(\ell)}\, T^{\mu(\ell)} \right) \right\rangle = \prod_{\ell=0}^{tN} \left[\frac{1}{p} \sum_{\nu=1}^{p} \exp\left(-\frac{i\eta\hat{x}}{N} \sigma^{tN-\ell} \boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\nu\, T^\nu \right) \right]$$

After substitution of this result into (25), only a training set average remains. Once this has been carried out, all terms in the sum over μ will be exactly equal. Anticipating this by setting $\mu = 1$, we get

$$\begin{aligned} \hat{P}_t(\hat{x}, \hat{y}, \hat{T}) &= \lim_{N\to\infty} \left\langle \exp[-i(\hat{x}\sigma^{tN} \boldsymbol{J}_0 \cdot \boldsymbol{\xi}^1 + \hat{y}\boldsymbol{B}^* \cdot \boldsymbol{\xi}^1 + \hat{T}T^1)] \right. \\ & \left. \times \prod_{\ell=0}^{tN} \left[\frac{1}{p} \sum_{\nu=1}^{p} \exp\left(-\frac{i\eta\hat{x}}{N} \sigma^{tN-\ell} \boldsymbol{\xi}^1 \cdot \boldsymbol{\xi}^\nu\, T^\nu \right) \right] \right\rangle_{\text{sets}} . \end{aligned} \tag{26}$$

Consider now the product $S = \prod_{\ell=0}^{tN}[\ldots]$. The $\nu = 1$ term of the sum in square brackets needs to be treated separately because $\boldsymbol{\xi}^1 \cdot \boldsymbol{\xi}^1 = N$. For $\nu > 1$, on the other hand, the products $\boldsymbol{\xi}^1 \cdot \boldsymbol{\xi}^\nu$

are overlaps between *different* input vectors and therefore only of $\mathcal{O}(\sqrt{N})$; the rescaled overlaps $v_\nu = \boldsymbol{\xi}^1 \cdot \boldsymbol{\xi}^\nu / \sqrt{N}$ are of $\mathcal{O}(1)$. In the sum over $\nu > 1$ in

$$\log S = \sum_{\ell=0}^{tN} \log \left[\frac{1}{p} \exp\left(-i\eta\hat{x}\sigma^{tN-\ell}T^1\right) + \frac{1}{p} \sum_{\nu>1} \exp\left(-\frac{i\eta\hat{x}}{\sqrt{N}}\sigma^{tN-\ell}v_\nu T^\nu\right)\right]$$

the second exponential therefore has an argument of $\mathcal{O}(N^{-1/2})$ and can be Taylor expanded. Terms up to $\mathcal{O}(1/N)$ (i.e., up to second order) need to be retained because of the sum over the $\mathcal{O}(N)$ values of ℓ, and so

$$\begin{aligned} \log S &= \\ &= \sum_{\ell=0}^{tN} \log \left[\frac{1}{p} \exp\left(-i\eta\hat{x}\sigma^{tN-\ell}T^1\right) + \frac{p-1}{p} + \frac{1}{p} \sum_{\nu>1} \left(-\frac{i\eta\hat{x}}{\sqrt{N}}\sigma^{tN-\ell}v_\nu T^\nu - \frac{\eta^2\hat{x}^2}{2N}\sigma^{2tN-2\ell}v_\nu^2\right)\right] \\ &= \frac{1}{p}\sum_{\ell=0}^{tN} \left[\exp\left(-i\eta\hat{x}\sigma^{tN-\ell}T^1\right) - 1 - i\eta\hat{x}\sigma^{tN-\ell}\frac{1}{\sqrt{N}}\sum_{\nu>1} v_\nu T^\nu - \frac{1}{2}\eta^2\hat{x}^2\sigma^{2tN-2\ell}\frac{1}{N}\sum_{\nu>1} v_\nu^2\right] \end{aligned}$$

where contributions of $\mathcal{O}(N^{-1/2})$ have been discarded. Transforming the first sum over l into an integral over time (by considering appropriate Riemann sums), we then obtain

$$\log S = \chi(\hat{x}T^1) - \frac{i\eta\hat{x}u_1}{\gamma}(1-e^{-\gamma t}) - \frac{\eta^2\hat{x}^2u_2}{4\gamma}(1-e^{-2\gamma t}) \tag{27}$$

where

$$\chi(w) = \frac{1}{\alpha}\int_0^t ds\, \left\{\exp\left[-i\eta w e^{-\gamma(t-s)}\right] - 1\right\} \tag{28}$$

and

$$u_1 = \frac{1}{\alpha\sqrt{N}}\sum_{\nu>1} v_\nu T^\nu \qquad u_2 = \frac{1}{p}\sum_{\nu>1} v_\nu^2\ .$$

Further progress requires considering the statistics of the random variables u_1 and u_2. For $N \to \infty$, the v_ν are independent Gaussian variables with zero mean and unit variance. By the central limit theorem, u_2 therefore has fluctuations of $\mathcal{O}(N^{-1/2})$ and can be replaced by its average $\langle u_2 \rangle = 1$ in the thermodynamic limit. Similarly, because the products $v_\nu T^\nu$ are uncorrelated for different ν, u_1 becomes Gaussian in this limit. Using (12), its mean and variance can be calculated as

$$\begin{aligned} \langle u_1 \rangle &= \frac{1}{\alpha\sqrt{N}}\sum_{\nu>1}\langle v_\nu T^\nu\rangle = \frac{p-1}{\alpha N}\left\langle \boldsymbol{\xi}^1 \cdot \boldsymbol{\xi}\, T(\boldsymbol{\xi})\right\rangle = \rho\, \boldsymbol{B}^* \cdot \boldsymbol{\xi}^1 + \mathcal{O}(N^{-1}) \\ \left\langle (\Delta u_1)^2 \right\rangle &= \frac{p-1}{\alpha^2 N}\left[\left\langle v_\nu^2\right\rangle - \langle v_\nu\rangle^2\right] = \frac{1}{\alpha}\left[1 - \mathcal{O}(N^{-1})\right] \end{aligned}$$

We conclude that, for large N, $u_1 = \rho\, \boldsymbol{B}^* \cdot \boldsymbol{\xi}^1 + \alpha^{-1/2}\hat{u}$, where $\hat{u}$ is a unit variance Gaussian random variable with mean zero. We are now in a position to average S as given by (27) over all realizations of $\{(\boldsymbol{\xi}^\nu, T^\nu), \nu > 1\}$, with the result

$$\langle S \rangle = \exp\left[\chi(\hat{x}T^1) - \frac{i\eta\hat{x}\rho\boldsymbol{B}^* \cdot \boldsymbol{\xi}^1}{\gamma}(1-e^{-\gamma t}) - \frac{\eta^2\hat{x}^2}{2\alpha\gamma^2}(1-e^{-\gamma t})^2 - \frac{\eta^2\hat{x}^2}{4\gamma}(1-e^{-2\gamma t})\right]$$

Inserting this into equation (26) for the characteristic function, we are left with a final average over $\boldsymbol{\xi}^1$ and T^1, with the former entering only through the fields $u = \boldsymbol{J}_0 \cdot \boldsymbol{\xi}^1$ and $y^1 = \boldsymbol{B}^* \cdot \boldsymbol{\xi}^1$:

$$\begin{aligned} \hat{P}_t(\hat{x}, \hat{y}, \hat{T}) \quad = \quad & \Bigg\langle \exp\Bigg[-i(\hat{x}e^{-\gamma t}u + \hat{y}y^1 + \hat{T}T^1) + \chi(\hat{x}T^1) - \frac{i\eta\hat{x}\rho y^1}{\gamma}(1 - e^{-\gamma t}) \\ & -\frac{\eta^2\hat{x}^2}{2\alpha\gamma^2}(1 - e^{-\gamma t})^2 - \frac{\eta^2\hat{x}^2}{4\gamma}(1 - e^{-2\gamma t}) \Bigg] \Bigg\rangle_{u, y^1, T^1} \end{aligned} \tag{29}$$

We now observe that T^1 only depends on y^1, but not on u; correspondingly, u is independent of T^1 if y^1 is given. For large N, the two fields u and y^1 are zero mean Gaussian random variables with $\langle u^2 \rangle = Q_0$, $\langle uy^1 \rangle = R_0$ and $\langle (y^1)^2 \rangle = 1$. The average of the u-dependent factor in (29), for given y^1, is therefore

$$\left\langle \exp\left(-i\hat{x}e^{-\gamma t}u \right) \right\rangle_{u|y^1} = \exp\left[-i\hat{x}e^{-\gamma t}R_0 y^1 - \frac{1}{2}\hat{x}^2 e^{-2\gamma t}(Q_0 - R_0^2) \right]$$

Inserting this into (29), and using (16,20), one finds that the terms in the exponential which are linear in $\hat{x}$ combine to a term proportional to $R(t)$, whereas the quadratic terms in $\hat{x}$ conspire to give a contribution proportional to $Q(t) - R^2(t)$:

$$\hat{P}_t(\hat{x}, \hat{y}, \hat{T}) = \left\langle \exp\left[-i\hat{y}y^1 - i\hat{T}T^1 + \chi(\hat{x}T^1) - \frac{1}{2}\hat{x}^2(Q - R^2) - iR\hat{x}y^1 \right] \right\rangle_{y^1, T^1} \tag{30}$$

Finally, we recast this result in terms of the conditional distribution of x, given y and T. To do this, first note that the distribution of y^1 and T^1 that is to be averaged over on the right hand side of (30) is just the distribution of the teacher field y and the teacher output T over the training set. We rename them appropriately and write out the definition (24) of the characteristic function on the left hand side:

$$\begin{aligned} & \int dx\, dy \sum_{T=\pm 1} \exp\left[-i\hat{y}y - i\hat{T}T - i\hat{x}x \right]\, P_t(x|y, T)\, P(y, T) = \\ & \qquad \int dy \sum_{T=\pm 1} \exp\left[-i\hat{y}y - i\hat{T}T + \chi(\hat{x}T) - \frac{1}{2}\hat{x}^2(Q - R^2) - iR\hat{x}y \right]\, P(y, T). \end{aligned}$$

Equality for all $\hat{y}$ and $\hat{T}$ implies that

$$\int dx\, \exp(-i\hat{x}x)\, P_t(x|y, T) = \exp\left[\chi(\hat{x}T) - \frac{1}{2}\hat{x}^2(Q - R^2) - iR\hat{x}y \right]$$

and hence our final result[3]

$$P_t(x|y, T) = \int \frac{d\hat{x}}{2\pi} \exp\left[i\hat{x}(x - Ry) + \chi(\hat{x}T) - \frac{1}{2}\hat{x}^2(Q - R^2) \right] \tag{31}$$

which is remarkably simple. In particular, we note that in this *conditional* distribution of x, the noise properties enter only through the parameter ρ; in fact, they only affect the factor

[3]Equation (31) can also be derived by using Fourier transforms to obtain $P_t(x, y, T)$ from (30), and then dividing by $P(y, T)$.

$\exp(-i\hat{x}Ry)$, while both $Q - R^2$ and $\chi(\hat{x}T)$ are actually independent of ρ. Eq. (31) also shows that the dependence of the student field on y and T can be written in the simple form

$$x = Ry + \Delta_1 + T\Delta_2$$

where Δ_1 and Δ_2 are random variables which are independent of each other and of y and T. Remarkably, they also do not depend on any properties of the noisy perceptron teacher: Δ_1 is simply Gaussian with zero mean and variance $Q - R^2$, while the distribution of Δ_2 follows from the characteristic function $\left\langle \exp(-i\hat{\Delta}\Delta_2) \right\rangle = \exp(\chi(\hat{\Delta}))$. All non-Gaussian features of the student field distribution are encoded in Δ_2. Because $\chi(\cdot)$ is inversely proportional to α, the size of the training set, it is immediately obvious how the student field distribution recovers its Gaussian form for $\alpha \to \infty$. More precisely, it can be shown that even for finite α, non-Gaussian effects in the distribution of x are negligible whenever $t \ll \alpha$. As before, this corresponds to the condition that training examples have not yet been 'recycled' in the learning process. Similarly, if $\alpha \gg 1/\gamma$, then the system reaches its asymptotic limit before it 'realizes' that α is finite; the distribution of x is then Gaussian for all times t. Finally, we note that in the absence of weight decay ($\gamma = 0$), the distribution of Δ_2 can be determined explicitly: Δ_2/η then obeys a Poisson distribution with mean t/α. By comparison with the explicit learning rule (11), it is easy to interpret this result for Δ_2. It represents the 'anomalous' contribution to the overlap $x = \boldsymbol{J} \cdot \boldsymbol{\xi}^\mu$ from the learning steps where the same example $\boldsymbol{\xi}^\mu$ was actually used to update $\boldsymbol{J}$; the number of such learning steps obviously has the required Poisson distribution. For nonzero weight decay, the distribution of Δ_2 broadens around the discrete values 0, η, 2η,. . . because the effect of each update with the same example is more or less heavily damped by the weight decay depending on when it took place.

Using the fact that y is Gaussian with zero mean and unit variance, the training error $E_{\rm tr}$ and student field probability density $P_t(x)$ follow from (31) as

$$E_{\rm tr} = \int dx\, Dy \sum_{T=\pm 1} \theta(-xT) P_t(x|y,T) P(T|y) \tag{32}$$

$$P_t(x) = \int Dy \sum_{T=\pm 1} P_t(x|y,T) P(T|y) \tag{33}$$

in which $Dy = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}y^2} dy$. We note again that the dependence of $E_{\rm tr}$ and $P_t(x)$ on the specific noise model – for a given value of ρ – arises solely through $P(T|y)$. We remind the reader that this teacher output probability is given by (3),

$$P(T|y) = (1 - \lambda)\ \theta(Ty) + \lambda\theta(-Ty)$$

for the case of output noise, while for weight noise (5) implies

$$P(T|y) = \frac{1}{2}[1 + T\,\mathrm{erf}(\mathrm{y}/\sqrt{2}\Sigma)].$$

In the appendix, we give explicit expressions for the training error and student field distribution in these two cases [see eqs. (48) to (51)], which also reveal a close relation between them.

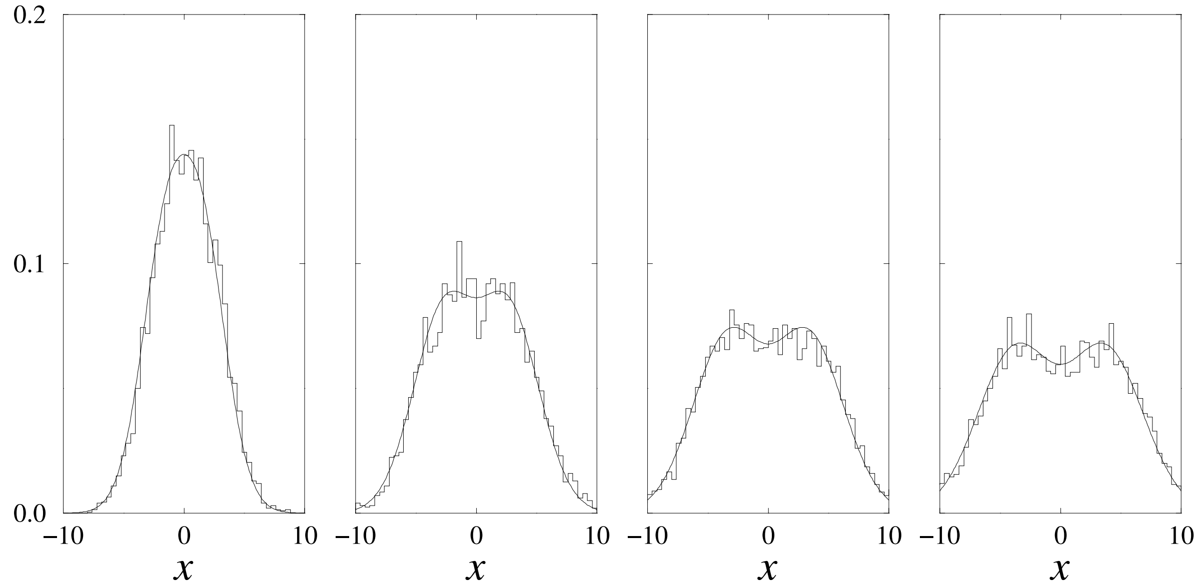

Figure 2: Student field distribution $P_t(x)$ observed during on-line Hebbian learning with output noise of strength $\lambda = 0.2$, at different times (from left to right: $t = 1, 2, 3, 4$), for training set size $\alpha = \frac{1}{2}$, learning rate $\eta = 1$, and weight decay $\gamma = \frac{1}{2}$, with initial conditions $Q_0 = 1$ and $R_0 = 0$. Histograms: distributions as measured in numerical simulations of an $N = 10{,}000$ system. Solid lines: predictions of the theory.

5 Comparison with Numerical Simulations

From the theoretical point of view, equations (31,32,33) constitute the clearest expression of our results on the joint field distribution since the dependence of the distribution on the given noise has been separated out in a transparent manner. However, we have found that another equivalent formulation can be useful from the point of view of numerical computations; this is detailed in the appendix.

It will be clear that there is a large number of parameters that one could vary in order to generate different simulation experiments with which to test our theory. Here we have to restrict ourselves to presenting a number of representative results. Figure 2 shows, for the output noise model, how the probability density $P_t(x)$ of the student fields $x = \boldsymbol{J} \cdot \boldsymbol{\xi}$ develops in time, starting as a Gaussian distribution at $t = 0$ (following random initialization of the student weight vector) and evolving into a highly non-Gaussian bi-modal one. Figure 3 compares our predictions for the generalization and training errors $E_{\rm g}$ and $E_{\rm tr}$ with the results of numerical simulations (again for teachers corrupted by output noise) for different initial conditions, $E_{\rm g,0} = 0$ and $E_{\rm g,0} = 0.5$, and for different choices of the two most important parameters λ (which controls the amount of teacher noise) and α (which measures the relative size of the training set). Different choices of the weight decay γ have also been explored, and yield similar results. The system is found to have no persistent memory of its past (which will be different for some other learning rules), the asymptotic values of $E_{\rm g}$ and $E_{\rm tr}$ being independent of the initial student vector[4].

[4]In the examples shown, $E_{\rm g}$ is always larger than $E_{\rm tr}$. However, this is not true generally: We are measuring the generalization error $E_{\rm g}$ with respect to the *clean* teacher, whereas the (training) examples that determine

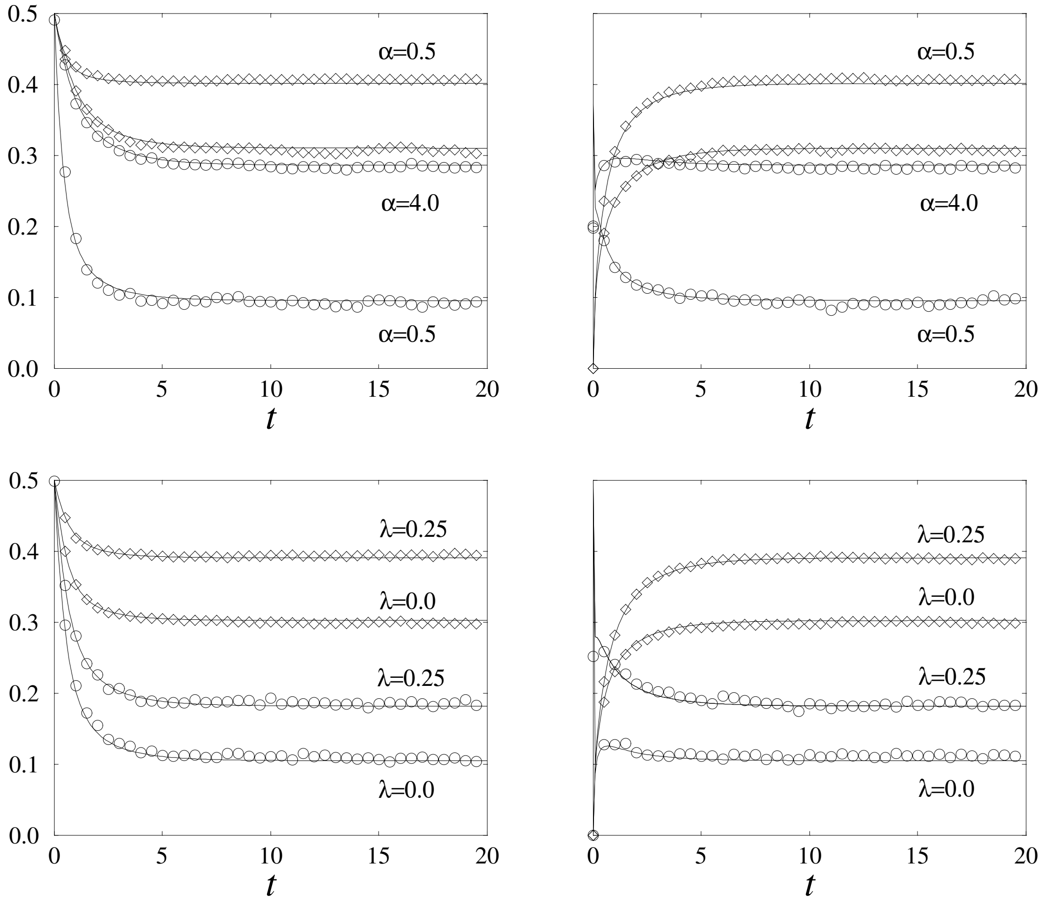

Figure 3: Generalization errors (diamonds/lines) and training errors (circles/lines) as observed during on-line Hebbian learning from a teacher corrupted by output noise, as functions of time. Upper two graphs: noise level $\lambda = 0.2$ and training set size $\alpha \in \{0.5, 4.0\}$ (initial conditions: upper left, $E_{\mathrm{g},0} = 0.5$; upper right: $E_{\mathrm{g},0} = 0$). Lower two graphs: $\alpha = 1$ and $\lambda \in \{0.0, 0.25\}$ (lower left, $E_{\mathrm{g},0} = 0.5$; lower right, $E_{\mathrm{g},0} = 0$). Markers: simulation results for an $N = 5{,}000$ system. Solid lines: predictions of the theory. In all cases $Q_0 = 1$, learning rate $\eta = 1$ and weight decay $\gamma = 0.5$.

Figure 4 shows the probability density $P_t(x)$ of the student fields $x = \boldsymbol{J} \cdot \boldsymbol{\xi}$ for the Gaussian weight noise model, with effective error probability λ_{eff} chosen identical to the error probability used to produce the corresponding graphs in figure 2 for output noise. Finally we show in figure 5 an example of a comparison between the error measures corresponding to teachers corrupted by output noise and teachers corrupted by Gaussian weight noise, both with identical effective output noise probability $\lambda_{\mathrm{eff}} = 0.2$. Here our theory predicts both noise types to exhibit identical

the training error E_{t} are *noisy*. Thus, under certain circumstances, E_{t} can be larger than E_{g}. A trivial example is the case of an infinite training set ($\alpha \to \infty$) without weight decay ($\gamma = 0$). From (17), E_{g} then tends to zero for long times t, while the training error will approach $E_{\mathrm{t}} = \lambda_{\mathrm{real}}$, which is nonzero for a noisy teacher. A generalization error relative to the noisy teacher can also be defined in our problem; it turns out to be $E_{\mathrm{g}}(\text{noisy}) = \{1 - \langle \tau(y)\,\mathrm{erf}(yR[2(Q - R^2)]^{-1/2})\rangle\}/2$.

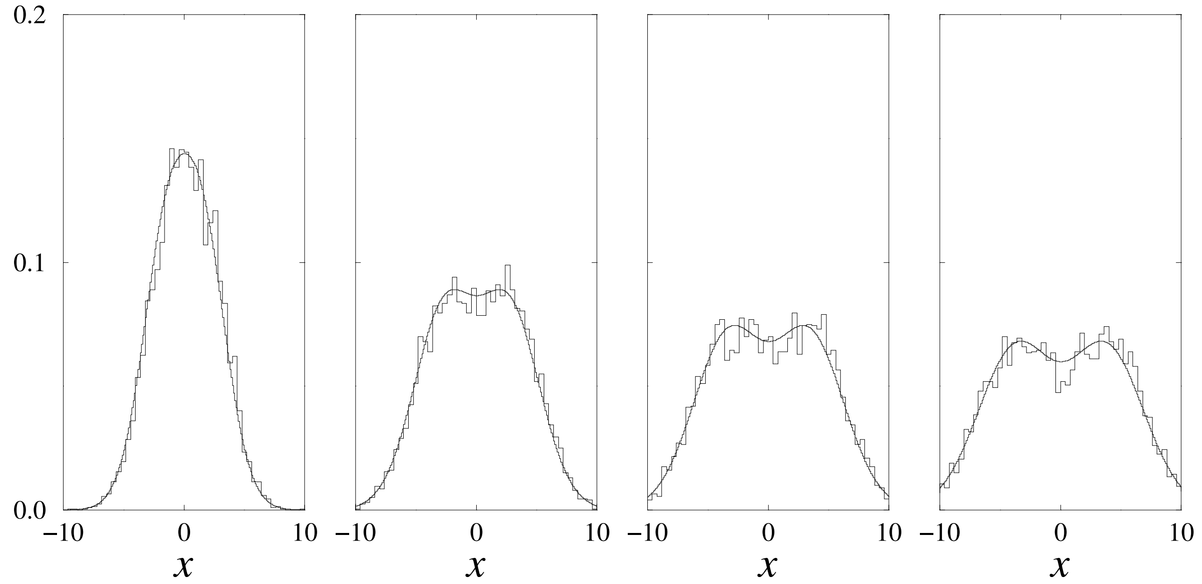

Figure 4: Student field distribution $P_t(x)$ observed during on-line Hebbian learning with Gaussian weight noise of effective error probability $\lambda_{\text{eff}} = 0.2$ (compare eq. (18)), at different times (from left to right: $t = 1, 2, 3, 4$), for training set size $\alpha = \frac{1}{2}$, learning rate $\eta = 1$, and weight decay $\gamma = \frac{1}{2}$, with initial conditions $Q_0 = 1$ and $R_0 = 0$. Histograms: distributions as measured in numerical simulations of an $N = 10{,}000$ system. Solid lines: predictions of the theory. See appendix for further discussion of the close similarities with figure 2.

generalization errors and almost identical training errors (with a difference of the order of 10^{-4}, see the appendix) at any time. These predictions are borne out by the corresponding numerical simulations (carried out with networks of size $N = 10{,}000$). We conclude from these figures that in all cases investigated the theoretical results give an extremely satisfactory account of the numerical simulations, with finite size effects being unimportant for the system sizes considered.

As pointed out in the theoretical analysis at the end of Section 3, there are no genuine overfitting effects in Hebbian learning with constant learning rate η. Any minima or maxima in $E_{\text{g}}(t)$ are due to the competition between forgetting a better-than-random initial generalization performance and learning a new set of weights with a different performance from the training data. We have run a number of simulations to address this point, and found our theoretical prediction confirmed. For time-dependent learning rates, on the other hand, preliminary theoretical work indicates that genuine overfitting effects can occur quite generically.

6 Conclusion

Starting from a microscopic description of Hebbian on-line learning in perceptrons with restricted training sets, of size $p = \alpha N$ where N is the number of inputs, we have developed an exact theory in terms of macroscopic observables which has enabled us to predict the generalization error and the training error, as well as the probability density of the student local fields, in the thermodynamic limit $N \to \infty$. Our results are found to be in excellent agreement with numerical simulations, as carried out for systems of size $N = 5{,}000$ and $N = 10{,}000$, and for various choices

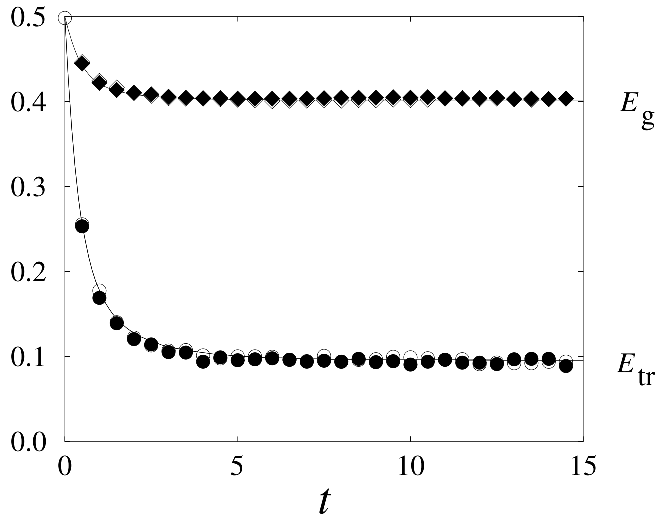

Figure 5: Comparison between output noise and Gaussian weight noise, with parameters such that both cases have identical effective error probability $\lambda_{\rm eff} = 0.2$. Open diamonds (output noise) and filled diamonds (weight noise): generalization errors as observed in numerical simulations, as functions of time. Open circles (output noise) and filled circles (weight noise): training errors as observed in numerical simulations, as functions of time. In all cases training set size $\alpha = 0.5$, learning rate $\eta = 1$, weight decay $\gamma = 0.5$, initial conditions $Q_0 = 1$ and $E_{g,0} = 0.5$, and system size $N = 10{,}000$. Solid lines: theory (which here predicts identical generalization errors and virtually identical training errors).

of the model parameters, both for teachers corrupted by output noise and for teachers corrupted by Gaussian input noise. Generalizations of our calculations to scenarios involving, for instance, time-dependent learning rates or time-dependent decay rates are straightforward. Closer analysis of the results for these cases, and for more complicated teachers such as noisy 'reversed wedges', may be an issue for future work.

Although it will be clear that our present calculations cannot be extended to non-Hebbian rules, since they ultimately rely on our ability to write down the microscopic weight vector $\boldsymbol{J}$ at any time in explicit form (11), they do indeed provide a significant yardstick against which more sophisticated and more general theories can be tested. In particular, they have already played a valuable role in assessing the conditions under which a recent general theory of learning with restricted training sets, based on a dynamical version of the replica formalism, is exact [10, 11].

Acknowledgments: PS is grateful to the Royal Society for financial support through a Dorothy Hodgkin Research Fellowship.

References

[1] Heskes T. and Kappen B. (1991). *Phys. Rev. A*, **44**, 2718.

[2] Biehl M. and Schwarze H. (1995). *J. Phys. A*, **28**, 643.

[3] Saad D. and Solla S. A. (1995). *Phys. Rev. E*, **52**, 4225.

[4] Saad D. (ed.), *On-line learning in neural networks*, Cambridge University Press, 1998.

[5] Mace C.W.H. and Coolen A.C.C. (1998) *Statistics and Computing* **8**, 55.

[6] Horner H. (1992), *Z. Phys. B* **86**, 291, and *Z. Phys. B* **87**, 371.

[7] Krogh A. and Hertz J.A. (1992) *J. Phys. A* **25**, 1135.

[8] Sollich P. and Barber D. (1997) *Europhys. Lett.* **38**, 477.

[9] Sollich P. and Barber D. (1998) *Advances in Neural Information Processing Systems 10*, Eds. Jordan M., Kearns M. and Solla S. A. (Cambridge: MIT), 357–363.

[10] Coolen A.C.C. and Saad D., in ref. [4], pages 303–343.

[11] Coolen A.C.C. and Saad D. (in preparation).

[12] Rae H.C., Sollich P. and Coolen A.C.C., in *Advances in Neural Information Processing Systems 10* (in press).

[13] Gradshteyn I.S. and Ryzhik I.M., *Tables of Integrals, Series, and Products* (*corrected and enlarged edition*), Academic Press, 1980.

[14] Titchmarsh E.C., *The Theory of Functions*, Oxford University Press, 1939.

A Evaluation of the Field Distribution and Training Error

In this appendix, we give alternative forms of our main results (31,32,33) for the joint field distribution and training error that are more suitable for numerical work. For this purpose, it is useful to shift attention from the noisy teacher output T to the corrupted teacher field z that produces it; the two are linked by $T = \mathrm{sgn}(z)$. This is entirely natural in the case of Gaussian weight noise. As discussed after eq. (4), z then differs from the clean teacher field y by an independent zero mean Gaussian variable with variance Σ^2; explicitly, one has the conditional distribution

$$P(z|y) = \frac{1}{\sqrt{2\pi\Sigma^2}}\, e^{-(z-y)^2/2\Sigma^2} \qquad \text{(Gaussian weight noise).}$$

The case of output noise can be treated similarly, by assuming that z is identical to y with probability $1-\lambda$, but has the opposite sign with probability λ:

$$P(z|y) = (1-\lambda)\,\delta(z-y) + \lambda\,\delta(z+y) \qquad \text{(output noise).} \tag{34}$$

We now consider the joint distribution $P_t(x, y, z)$. It can be derived by complete analogy with the calculation in section 4. For the conditional distribution of x, one finds that

$$P_t(x|y, z) = P_t(x|y, \mathrm{sgn}(z)).$$

Intuitively, this follows from the fact that during learning, the student only ever sees the noisy teacher output $\mathrm{sgn}(z)$, but not the corrupted field z itself; the student field x can therefore depend on z only through $\mathrm{sgn}(z)$. Multiplying by the joint distribution of y and z, and using the result (31), one thus finds, for the case of output noise,

$$P_t(x, y, z) = [(1-\lambda)\,\delta(z-y) + \lambda\,\delta(z+y)]\, \frac{e^{-\frac{1}{2}y^2}}{\sqrt{2\pi}} \int \frac{d\hat{x}}{2\pi}\, e^{-\frac{1}{2}\hat{x}^2(Q-R^2)+i\hat{x}(x-yR)+\chi(\hat{x}\mathrm{sgn}(z))}$$

with the marginal distribution

$$P_t(x, z) = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} \int \frac{d\hat{x}}{2\pi}\, e^{-\frac{1}{2}\hat{x}^2(Q-R^2)+i\hat{x}x+\chi(\hat{x}\mathrm{sgn}(z))} \left[(1-\lambda)e^{-i\hat{x}zR} + \lambda e^{i\hat{x}zR}\right]. \tag{35}$$

The corresponding expressions in the case of Gaussian weight noise read

$$P_t(x, y, z) = \frac{1}{2\pi\Sigma} \int \frac{d\hat{x}}{2\pi}\, e^{-\frac{1}{2}\hat{x}^2(Q-R^2)+i\hat{x}(x-Ry)+\chi(\hat{x}\mathrm{sgn}(z))-[z^2-2yz+y^2(1+\Sigma^2)]/(2\Sigma^2)}$$

and

$$P_t(x, z) = \frac{e^{-\frac{1}{2}z^2/(1+\Sigma^2)}}{\sqrt{2\pi(1+\Sigma^2)}} \int \frac{d\hat{x}}{2\pi}\, e^{-\frac{1}{2}\hat{x}^2[Q-R^2/(1+\Sigma^2)]+i\hat{x}[x-Rz/(1+\Sigma^2)]+\chi(\hat{x}\mathrm{sgn}(z))}\,. \tag{36}$$

In both cases, the training error and the probability distribution of the student field x are then determined by

$$E_{\mathrm{tr}} = \int dx\, dz\, \theta(-xz)\, P_t(x, z) \qquad P_t(x) = \int dz\, P_t(x, z)$$

respectively. For a numerical computation of these two quantities, it is imperative to further reduce the number of integrations analytically, which turns out to be possible. In the following, we drop the time subscript t on all distributions to save notation.

First we deal with the case of output noise. In the marginal distribution (35), we make the change of variable $\hat{x} = k\,\mathrm{sgn}(z)$ to get

$$P(x,z) = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} \int \frac{dk}{2\pi}\, e^{-\frac{1}{2}k^2(Q-R^2)+\chi(k)+ikx\mathrm{sgn}(z)} \left\{(1-\lambda)e^{-ik|z|R} + \lambda e^{ik|z|R}\right\}.$$

The training error is

$$E_{\mathrm{tr}} = \int dx\, dz\, P(x,z)\, \theta(-xz) = \int_0^\infty dx\ [P_+(-x) + P_-(x)]$$

where

$$P_\pm(x) = \int dz\, P(x,z)\, \theta(\pm z) = \frac{1}{2}\int \frac{dk}{2\pi}\, Dz\, e^{-\frac{1}{2}k^2(Q-R^2)+\chi(k)\pm ikx} \left\{(1-\lambda)e^{-ik|z|R} + \lambda e^{ik|z|R}\right\} \tag{37}$$

We see that $P_+(x) = P_-(-x) \equiv \Pi(x)$. In terms of $\Pi(x)$ we have the formulae

$$P(x) = \Pi(x) + \Pi(-x) \qquad E_{\mathrm{tr}} = 2\int_0^\infty dx\, \Pi(-x) \tag{38}$$

The function $\Pi(x)$ can be further simplified by decomposing χ into its real ($\chi_\mathrm{r} = \mathrm{Re}(\chi)$) and imaginary ($\chi_\mathrm{i} = \mathrm{Im}(\chi)$) parts:

$$\begin{aligned}
\Pi(x) &= \int \frac{dk}{4\pi}\, Dz\, e^{-\frac{1}{2}k^2(Q-R^2)+\chi(k)+ikx} \left\{(1-\lambda)e^{-ik|z|R} + \lambda e^{ik|z|R}\right\} \\
&= \int \frac{dk}{4\pi}\, Dz\, e^{-\frac{1}{2}k^2(Q-R^2)+\chi_\mathrm{r}(k)} \left\{(1-\lambda)\cos[\chi_\mathrm{i}(k) + k(x - R|z|)] + \lambda\cos[\chi_\mathrm{i}(k) + k(x + R|z|)]\right\} \\
&= \int \frac{dk}{4\pi}\, e^{-\frac{1}{2}Qk^2+\chi_\mathrm{r}(k)} \left\{\cos[\chi_\mathrm{i}(k) + kx] + (1-2\lambda)\sin[\chi_\mathrm{i}(k) + kx]\, G(kR)\right\}
\end{aligned} \tag{39}$$

in which

$$G(\Lambda) = e^{\frac{1}{2}\Lambda^2}\int Dz\ \sin(\Lambda|z|) = \frac{\Lambda}{\sqrt{\pi}}\ {}_1F_1\left(\frac{1}{2};\frac{3}{2};\frac{1}{2}\Lambda^2\right) \tag{40}$$

and ${}_1F_1(\ldots)$ is the degenerate hypergeometric function (see [13], page 1058). From equation (38) we now immediately obtain our final result for the student field distribution:

$$P(x) = \int \frac{dk}{2\pi}\, e^{-\frac{1}{2}Qk^2+\chi_\mathrm{r}(k)} \cos(kx) \left\{\cos[\chi_\mathrm{i}(k)] + (1-2\lambda)G(kR)\sin[\chi_\mathrm{i}(k)]\right\} \tag{41}$$

To further simplify the expression (38) for the training error, we write

$$E_{\mathrm{tr}} = \lim_{L\to\infty} 2\int_{-L}^{0} dx\, \Pi(x) = 2\lim_{L\to\infty} I(L)$$

where, from (39)

$$I(L) \;=\; \int \frac{dk}{4\pi}\, e^{-\frac{1}{2}Qk^2+\chi_\mathrm{r}(k)} \left\{\int_{-L}^{0} dx\ \cos[\chi_\mathrm{i}(k) + kx] + (1-2\lambda)G(kR)\int_{-L}^{0} dx\ \sin[\chi_\mathrm{i}(k) + kx]\right\}$$

Thus

$$\begin{aligned}
I(\infty) \;=\; &-\int \frac{dk}{4\pi k}\, e^{-\frac{1}{2}Qk^2+\chi_\mathrm{r}(k)} \left\{(1-2\lambda)G(kR)\cos[\chi_\mathrm{i}(k)] - \sin[\chi_\mathrm{i}(k)]\right\} + \\
&\lim_{L\to\infty}\int \frac{dk}{4\pi k}\, e^{-\frac{1}{2}Qk^2+\chi_\mathrm{r}(k)} \left\{\sin[kL - \chi_\mathrm{i}(k)] + (1-2\lambda)G(kR)\cos[kL - \chi_\mathrm{i}(k)]\right\}
\end{aligned} \tag{42}$$

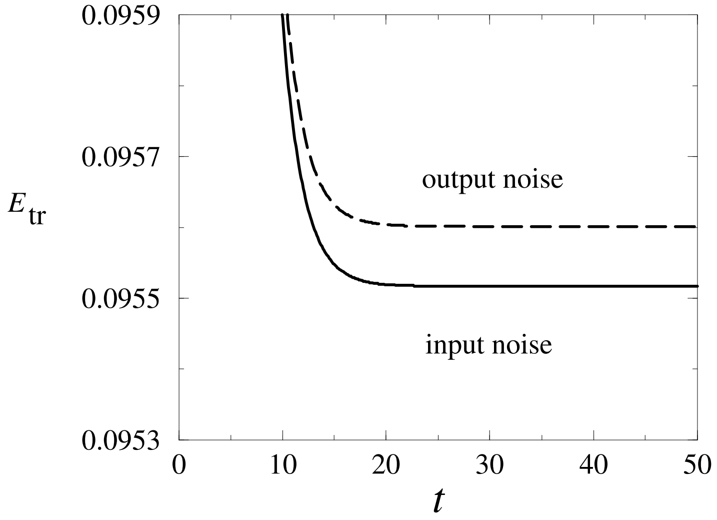

Figure 6: Characteristic example of theoretical predictions for the training error $E_{\rm tr}$ for two noisy teachers with identical effective error probability $\lambda_{\rm eff} = 0.2$. Dashed line: output noise; solid line: Gaussian weight noise. Parameters: $\alpha = \gamma = 0.5$, $Q_0 = \eta = 1$, $E_{\rm g,0} = 0.5$.

The L-dependent integral in (42) can be expressed as a sum of two integrals, which we consider separately. In the first part, we replace k by k/L and obtain

$$\lim_{L\to\infty} \int \frac{dk}{4\pi k}\, e^{-\frac{1}{2}Qk^2+\chi_{\rm r}(k)} \sin[kL - \chi_{\rm i}(k)]$$
$$= \lim_{L\to\infty} \int \frac{dk}{4\pi k}\, e^{-\frac{1}{2}Q(k/L)^2+\chi_{\rm r}(k/L)} \sin[k - \chi_{\rm i}(k/L)] = \int \frac{dk}{4\pi k}\, \sin(k) = \frac{1}{4}\ .$$

Secondly we need to consider the behaviour of

$$\int \frac{dk}{4\pi k}\, e^{-\frac{1}{2}Qk^2+\chi_{\rm r}(k)} \cos[kL - \chi_{\rm i}(k)] G(kR) \tag{43}$$

in the limit $L \to \infty$. We set $u = kR$ and note that, because $Q \geq R^2$, one has $e^{-\frac{1}{2}Qk^2} \leq e^{-\frac{1}{2}u^2}$; furthermore,

$$\left| e^{-\frac{1}{2}u^2} G(u) u^{-1} \right| = \left| \int Dz\, |z| \frac{\sin(|uz|)}{|uz|} \right| \leq \int Dz\, |z| = \sqrt{\frac{2}{\pi}}\ .$$

Finally, $\chi(k)$ is independent of L and is bounded as a function of k; in fact, from (28), $|\chi(k)| \leq 2\alpha^{-1}t$. It follows by an application of the Riemann-Lebesgue Lemma (see e.g. [14]) that the integral (43) tends to zero as $L \to \infty$. We conclude that for output noise the training error is given by

$$E_{\rm tr} = \frac{1}{2} - \int \frac{dk}{2\pi k}\, e^{-\frac{1}{2}Qk^2+\chi_{\rm r}(k)} \left\{ (1-2\lambda) G(kR) \cos[\chi_{\rm i}(k)] - \sin[\chi_{\rm i}(k)] \right\} \tag{44}$$

where $G(\ldots)$ is defined by (40).

The procedure for Gaussian weight noise is similar to that of output noise. We start from equation (36) and define

$$\widetilde{R} = R/\sqrt{1+\Sigma^2}.$$

Upon defining $\hat{x} = k\,\mathrm{sgn}(z)$ in (36), replacing z by $z/\sqrt{1+\Sigma^2}$, and continuing in the same notation as for output noise, we find

$$P_{\pm}(x) = \frac{1}{2}\int \frac{dk}{2\pi}\, Dz\, e^{-\frac{1}{2}k^2(Q-\widetilde{R}^2)+\chi(k)\pm ikx - ik\widetilde{R}|z|} \tag{45}$$

Since (45) can be obtained from (37) by putting $\lambda \to 0$ and $R \to \widetilde{R}$, we immediately obtain for the student field distribution and the training error, respectively [see equations (41) and (44)],

$$P(x) = \int \frac{dk}{2\pi}\, e^{-\frac{1}{2}Qk^2+\chi_{\rm r}(k)} \cos(kx) \left\{\cos[\chi_{\rm i}(k)] + G(k\widetilde{R})\sin[\chi_{\rm i}(k)]\right\} \tag{46}$$

$$E_{\rm tr} = \frac{1}{2} - \int \frac{dk}{2\pi k}\, e^{-\frac{1}{2}Qk^2+\chi_{\rm r}(k)} \left\{G(k\widetilde{R})\cos[\chi_{\rm i}(k)] - \sin[\chi_{\rm i}(k)]\right\} \tag{47}$$

In particular, we can now calculate the student field distribution and the training error for both output noise and Gaussian weight noise, with noise levels such that in both cases $\lambda_{\rm eff} = \lambda$. This guarantees that, at any time, Q, R and $E_{\rm g}$ will have the same values in both cases; it also implies $\widetilde{R} = R/\sqrt{1+\Sigma^2} = R(1-2\lambda)$. We then obtain from (41,44,46,47) very similar expressions:

$$P^{\rm out}(x) = \int \frac{dk}{2\pi}\, e^{-\frac{1}{2}Qk^2+\chi_{\rm r}(k)} \cos(kx) \left\{\cos[\chi_{\rm i}(k)] + (1-2\lambda)G(kR)\sin[\chi_{\rm i}(k)]\right\} \tag{48}$$

$$P^{\rm gau}(x) = \int \frac{dk}{2\pi}\, e^{-\frac{1}{2}Qk^2+\chi_{\rm r}(k)} \cos(kx) \left\{\cos[\chi_{\rm i}(k)] + G[(1-2\lambda)kR]\sin[\chi_{\rm i}(k)]\right\} \tag{49}$$

and

$$E_{\rm tr}^{\rm out} = \frac{1}{2} - \int \frac{dk}{2\pi k}\, e^{-\frac{1}{2}Qk^2+\chi_{\rm r}(k)} \left\{(1-2\lambda)G(kR)\cos[\chi_{\rm i}(k)] - \sin[\chi_{\rm i}(k)]\right\} \tag{50}$$

$$E_{\rm tr}^{\rm gau} = \frac{1}{2} - \int \frac{dk}{2\pi k}\, e^{-\frac{1}{2}Qk^2+\chi_{\rm r}(k)} \left\{G[(1-2\lambda)kR]\cos[\chi_{\rm i}(k)] - \sin[\chi_{\rm i}(k)]\right\} \tag{51}$$

Provided parameters are chosen such that the effective error probabilities are identical, the differences between output noise and Gaussian weight noise are restricted to the positioning of the factor $1-2\lambda$ relative to the integral $G(\ldots)$, with manifestly identical expressions for $\lambda = 0$ and $\lambda = \frac{1}{2}$ (as it should be). As a result the resulting curves for field distributions and training errors are found to be almost identical; figure 6 shows a typical example.