J. Phys. A: Math. Gen. **33** (2000) 8703–8722. Printed in the UK PII: S0305-4470(00)15429-0

Non-deterministic learning dynamics in large neural networks due to structural data bias

H C Rae, J A F Heimel and A C C Coolen

Department of Mathematics, King's College London, The Strand, London WC2R 2LS, UK

Received 13 July 2000

Abstract. We study the dynamics of on-line learning in large ($N \to \infty$) perceptrons, for the case of training sets with a structural $\mathcal{O}(N^0)$ bias of the input vectors, by deriving exact and closed macroscopic dynamical laws using non-equilibrium statistical mechanical tools. In sharp contrast to the more conventional theories developed for homogeneously distributed or only weakly biased data, these laws are found to describe a non-trivial and persistently non-deterministic macroscopic evolution, and a generalization error which retains both stochastic and sample-to-sample fluctuations, even for infinitely large networks. Furthermore, for the standard error-correcting microscopic algorithms (such as the perceptron learning rule) one obtains learning curves with distinct bias-induced phases. Our theoretical predictions find excellent confirmation in numerical simulations.

1. Introduction

Rosenblatt [1] first introduced the perceptron and proved the famous perceptron convergence theorem in 1962. It is an indicator of the richness of the perceptron as a dynamical system that almost 40 years later it continues to yield fascinating results which have hitherto remained hidden. Especially during the last decade, considerable progress has been made in understanding the dynamics of learning in artificial neural networks through the application of the methods of statistical mechanics. The dynamics of on-line learning in perceptrons has been analysed intensively, but for the most part such studies [2] have been carried out in the idealized scenario of so-called complete training sets (in which the number of training examples is large compared with N, the number of degrees of freedom), and have also assumed a homogeneous input data distribution. A recent review of work in this field is contained in [3]. A general theory of learning in the context of restricted training sets (where the size of the training set is proportional to N) is generally much more difficult, although an exact solution of the dynamical equations for the more elementary problem of unbiased on-line Hebbian learning with restricted training sets and noisy teachers has been found [4, 5]. Nevertheless, substantial progress has been made towards a general theory of the dynamics of on-line learning with restricted training sets and the reader may refer, for example, to [6–9], for details.

In this paper we consider complete training sets, but we admit the possibility of a structural bias of the input vectors. This is a significant issue since in real-world situations a training sample will generally have a non-zero average; this is especially important in the case of on-line learning, where examples are not available prior to learning, so that one cannot correct for any bias prior to processing. This in itself would be sufficient motivation for the present study. However, it turns out that the introduction of structurally biased input data leads to

0305-4470/00/488703+20$30.00

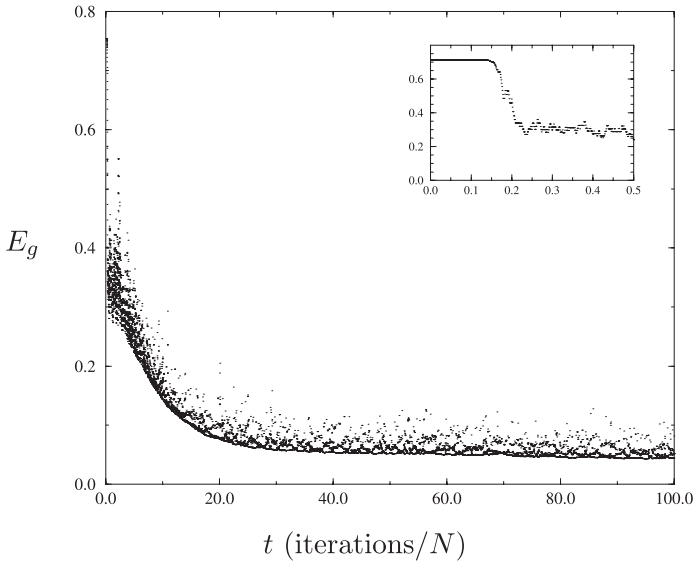

Figure 1. Evolution of the generalization error E_g as measured in a single simulation experiment of the perceptron rule, with $N = 1000$, learning rate $\eta = 1$ and bias $a = \frac{1}{2}$, following initial conditions $Q(0) = 10$, $R(0) = 0$ and $S(0) = \sqrt{N}$ (see the text for details). The inset, magnifying the early transients, shows phases I and II. Clearly, no learning takes place in phase I.

qualitative (rather than only quantitative) modifications of the actual learning curves observed in numerical simulations and the mathematical theories required for their description. Various authors [10, 11] have studied so-called clustered examples, in which examples are drawn from two Gaussian distributions situated close to each other, with an input bias of the order of $N^{-1/2}$ (i.e. in magnitude similar to finite-size effects). Learning with input bias has also been considered in the context of linear networks [12]; the linear theory was then used to construct an approximation for a class of nonlinear models, and it was shown that on-line learning is more robust to input bias and outperforms batch learning when such bias is present.

Here we consider a situation which is more natural and less restrictive than that considered in [10, 11], and which does not require the linearity of [12]: we study the familiar (nonlinear) perceptron, with the perceptron learning rule and with a structural, i.e. $\mathcal{O}(N^0)$, bias in the input data. Using $\boldsymbol{B}$, $\boldsymbol{J}$ and $\boldsymbol{A}$ to denote the teacher weights, the student weights and the bias vector (precise definitions follow), we develop our theory in terms of three macroscopic observables: the standard observables $Q = \boldsymbol{J}^2$, $R = \boldsymbol{J} \cdot \boldsymbol{B}$, and a new observable $S = \boldsymbol{J} \cdot \boldsymbol{A}$ (the overlap between student weights and the bias vector). In contrast to the dynamics of the bias-free case, we find that in the presence of an $\mathcal{O}(N^0)$ input bias the system passes though three phases, characterized by different scalings of typical times and of macroscopic observables. This could already have been anticipated on the basis of numerical simulations (see, e.g., figure 1). We obtain a closed system of equations in which the evolution of $\{Q, R\}$ is deterministic in the limit $N \to \infty$, as in the bias-free case, but where S is (generally) a stochastic variable, whose conditional probability distribution $P_t(S|Q, R)$ becomes non-trivial. Phase I is a short phase, in which the system reduces the alignment of the student weight vector $\boldsymbol{J}$ relative to the bias vector $\boldsymbol{A}$. During phase I, the observable S is deterministic, is rapidly driven towards zero and no learning takes place. Before the state $S = 0$ is reached, however, the system enters phase II, a very short phase in which S evolves stochastically to a quasi-stationary probability

distribution (which we calculate) and in which both Q and R are frozen. In phase III, where most of the learning takes place, the S distribution is modified by a non-negligible random walk element, which generates a diffusion term in the equation controlling the evolution of $P_t(S|Q, R)$, whereas Q and R satisfy coupled differential equations which involve averages over $P_t(S|Q, R)$. The stochastic nature of S is reflected in the fact that the generalization error also exhibits fluctuations (see figure 1). The (exact) equations describing phase III cannot be further simplified, but we introduce an approximation yielding more tractable equations for $\{Q, R\}$, which still have the merit of reducing to the more familiar equations when no-bias is present. Moreover, they are found to be in excellent agreement with the results obtained from numerical simulations. Compared with the unbiased case, having a finite bias is found to change the pre-factor in the asymptotic power law of the asymptotic time dependence of the generalization error, but not the exponent. A preliminary and more intuitive presentation of some of the present results can be found in [13].

2. Definitions

We study on-line learning in a student perceptron $\Sigma : \{-1, 1\}^N \to \{-1, 1\}$, which learns a task defined by a teacher perceptron $T : \{-1, 1\}^N \to \{-1, 1\}$ whose fixed weight vector is $\boldsymbol{B} \in \Re^N$. The teacher and student output are given by the familiar recipes

$$T(\boldsymbol{\xi}) = \operatorname{sgn}[\boldsymbol{B} \cdot \boldsymbol{\xi}] \qquad \Sigma(\boldsymbol{\xi}) = \operatorname{sgn}[\boldsymbol{J} \cdot \boldsymbol{\xi}].$$

We assume that $\boldsymbol{B}$ is normalized such that $\boldsymbol{B}^2 = 1$, the components being drawn randomly with mean zero and standard deviation of the order of $\mathcal{O}(N^{-1/2})$, and statistically independent of the input data. In order to model the bias in the input sample we assume that for $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N) \in \{-1, 1\}^N$ all ξ_i are independent, with $\langle \xi_i \rangle = a$, so that the probability of drawing $\boldsymbol{\xi}$ is given by

$$p(\boldsymbol{\xi}) = \prod_i \tfrac{1}{2}[1 + a\xi_i]. \tag{1}$$

We define $\xi_i = a + v_i$, such that the (independent) v_i have a mean of zero and variance $\sigma^2 = 1 - a^2$, and the shorthand $\boldsymbol{A} = a(1, \ldots, 1)$ (i.e. a vector with all N entries equal to a, to be referred to as the 'bias vector'). The teacher-bias overlap $\boldsymbol{B} \cdot \boldsymbol{A}$ is now a random parameter which is $\mathcal{O}(1)$, since $\langle (\boldsymbol{B} \cdot \boldsymbol{A})^2 \rangle = \sum_{i,j} a^2 \langle B_i B_j \rangle = a^2$, whose distribution will be Gaussian for $N \to \infty$, with mean 0 and standard deviation a.

The student perceptron Σ is being trained according to an on-line learning rule of the form $\boldsymbol{J}_{m+1} = \boldsymbol{J}_m + \Delta \boldsymbol{J}_m$, where at each iteration step an input vector $\boldsymbol{\xi}_m$ is drawn independently according to (1), and where

$$\Delta \boldsymbol{J}_m = \frac{\eta}{N} \boldsymbol{\xi}_m \operatorname{sgn}(\boldsymbol{B} \cdot \boldsymbol{\xi}_m) \mathcal{F}[|\boldsymbol{J}_m|, \boldsymbol{J}_m \cdot \boldsymbol{\xi}_m, \operatorname{sgn}(\boldsymbol{B} \cdot \boldsymbol{\xi}_m)].$$

For Hebbian learning, for instance, we have

$$\mathcal{F}[J, u, T] = 1: \qquad \Delta \boldsymbol{J}_m = \frac{\eta}{N} \boldsymbol{\xi}_m \operatorname{sgn}(\boldsymbol{B} \cdot \boldsymbol{\xi}_m)$$

whilst the familiar perceptron learning rule is defined by

$$\mathcal{F}[J, u, T] = \theta[-uT]: \qquad \Delta \boldsymbol{J}_m = \frac{\eta}{2N} \boldsymbol{\xi}_m [\operatorname{sgn}(\boldsymbol{B} \cdot \boldsymbol{\xi}_m) - \operatorname{sgn}(\boldsymbol{J}_m \cdot \boldsymbol{\xi}_m)]. \tag{2}$$

We will derive, from the microscopic stochastic process for the weight vector $\boldsymbol{J}$, a macroscopic dynamical theory in terms of the familiar observables $Q = \boldsymbol{J}^2$ and $R = \boldsymbol{J} \cdot \boldsymbol{B}$, as well as (in

order to obtain closure) a new observable $S = \boldsymbol{J} \cdot \boldsymbol{A}$ measuring the overlap between the vector $\boldsymbol{J}$ and the bias vector. The student and teacher output can then be written in the form

$$\Sigma(\boldsymbol{\xi}) = \mathrm{sgn}[\lambda_1 + x] \qquad T(\boldsymbol{\xi}) = \mathrm{sgn}[\lambda_2 + y] \qquad \text{with} \quad \lambda_1 = \hat{\boldsymbol{J}} \cdot \boldsymbol{A} \quad \lambda_2 = \boldsymbol{B} \cdot \boldsymbol{A}$$

with $\hat{\boldsymbol{J}} = \boldsymbol{J}/|\boldsymbol{J}|$, and where the local fields $\{x, y, z\}$ are defined by $x = \hat{\boldsymbol{J}} \cdot \boldsymbol{v}$, $y = \boldsymbol{B} \cdot \boldsymbol{v}$ and $z = \hat{\boldsymbol{A}} \cdot \boldsymbol{v}$ (the latter field z will also enter our calculation in due course). Note that $\lambda_1 = S/\sqrt{Q}$. For large N, the three fields $\{x, y, z\}$ are zero-average Gaussian random variables, each with variance $\sigma^2 = 1 - a^2$, and with correlation coefficients given by

$$\langle xy \rangle = \omega\sigma^2 \qquad \langle xz \rangle = \sigma^2 S/|\boldsymbol{A}| \qquad \langle yz \rangle = \sigma^2 \lambda_2/|\boldsymbol{A}|. \tag{3}$$

We note that equation (3) implies that z will be independent of (x, y) for large N if $S = \mathrm{o}(\sqrt{N})$ so that

$$\begin{aligned} p(x, y, z) &= \left[\sigma\sqrt{2\pi}\,\right]^{-1} \mathrm{e}^{-z^2/2\sigma^2}\, p(x, y) \\ p(x, y) &= \left[2\pi\sigma^2\sqrt{1-\omega^2}\,\right]^{-1} \mathrm{e}^{-\frac{1}{2}[x^2 - 2\omega xy + y^2]/\sigma^2(1-\omega^2)} \end{aligned} \tag{4}$$

with $\omega = \hat{\boldsymbol{J}} \cdot \boldsymbol{B} = R/\sqrt{Q}$. It will turn out that most of the averages to appear in this paper, involving (4) (to be written as $\langle \cdots \rangle$), may be expressed in terms of the function $K(x) = \mathrm{erf}\,(x/\sqrt{2})$. The generalization error $E_g = \langle \theta[-(\hat{\boldsymbol{J}} \cdot \boldsymbol{\xi})(\boldsymbol{B} \cdot \boldsymbol{\xi})] \rangle$, for example, can be written as

$$E_g = \int \mathrm{d}x\, \mathrm{d}y\; p(x, y)\theta[-(\lambda_1 + x)(\lambda_2 + y)] = I_1(\lambda_1, -\lambda_2, -\omega) + I_1(-\lambda_1, \lambda_2, -\omega) \tag{5}$$

where

$$I_1(\lambda_1, \lambda_2, \omega) = \int_{\lambda_1}^{\infty} \mathrm{d}x \int_{\lambda_2}^{\infty} \mathrm{d}y\; p(x, y) = \frac{1}{4}\left[1 - K\left(\frac{\lambda_2}{\sigma}\right)\right] - \frac{1}{2}\int_{\lambda_2/\sigma}^{\infty} Dy\; K\left(\frac{\lambda_1 - \omega\sigma y}{\sigma\sqrt{1-\omega^2}}\right)$$

with the Gaussian measure $Dy = (2\pi)^{-1/2}\mathrm{e}^{-\frac{1}{2}y^2}\mathrm{d}y$ (see appendix A for details). This then gives

$$E_g = \frac{1}{2} - \frac{1}{2}\int_{-\lambda_2/\sigma}^{\infty} Dy\; K\left(\frac{\lambda_1 + \omega\sigma y}{\sigma\sqrt{1-\omega^2}}\right) + \frac{1}{2}\int_{\lambda_2/\sigma}^{\infty} Dy\; K\left(\frac{\lambda_1 - \omega\sigma y}{\sigma\sqrt{1-\omega^2}}\right). \tag{6}$$

Note that, due to the identity $\int_0^{\infty} Dy\, K(\omega y/\sqrt{1-\omega^2}) = \frac{1}{2} - \frac{1}{\pi}\arccos\omega$, formula (6) reduces, as it should, to the well known expression $E_g = \pi^{-1}\arccos\omega$ in the case where the input bias is zero (i.e. for $a \to 0$).

3. From microscopic to macroscopic laws

We now consider the dynamics of the macroscopic observables $\{Q, R, S\}$ in the limit of large N. In the bias-free case, where for large N the fluctuations in the macroscopic observables are insignificant, this can be done in a direct and simple way. Here, for $a \neq 0$, the situation is qualitatively different, since (as it will turn out) the fluctuations in S will no longer vanish, and their distribution will have a strong impact on the macroscopic laws. In order to provide a setting for our theory we briefly review a well known procedure [3] which enables us to pass from a discrete to a continuous time description. We suppose that at time t the probability that the perceptron has undergone precisely m updates is given by the Poisson distribution $\pi_m(t) = \frac{1}{m!}(Nt)^m\mathrm{e}^{-Nt}$. For large N this will give us $t = \frac{m}{N} + \mathcal{O}(N^{-1/2})$, the usual real-valued time unit, and the uncertainty as to where we are on the time axis vanishes as $N \to \infty$. It is not hard to show that the probability density $p_t(\boldsymbol{J})$ of finding the vector $\boldsymbol{J}$ at time t satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\, p_t(\boldsymbol{J}) = N \int \mathrm{d}\boldsymbol{J}' \left\{ \langle \delta[\boldsymbol{J} - \boldsymbol{J}' - \Delta\boldsymbol{J}] \rangle_{\xi} - \delta[\boldsymbol{J} - \boldsymbol{J}'] \right\} p_t(\boldsymbol{J}')$$

where, for the perceptron learning rule (2), the single-step modification $\Delta \boldsymbol{J}$ is given by

$$\Delta \boldsymbol{J} = \frac{\eta}{2N} \boldsymbol{\xi} \left[\operatorname{sgn}(\boldsymbol{B} \cdot \boldsymbol{\xi}) - \operatorname{sgn}(\boldsymbol{J} \cdot \boldsymbol{\xi}) \right]$$

and where and $\langle \cdots \rangle_{\xi}$ denotes the average over all questions $\boldsymbol{\xi}$ in the training set $\{-1, 1\}^N$. The macroscopic observables $\boldsymbol{\Omega} = (Q, R, S)$, in turn, have the probability density $P_t(\boldsymbol{\Omega}) = \int \mathrm{d}\boldsymbol{J}\, p_t(\boldsymbol{J}) \delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}(\boldsymbol{J})]$, which satisfies the macroscopic stochastic equation

$$\frac{\mathrm{d}}{\mathrm{d}t} P_t(\boldsymbol{\Omega}) = \int \mathrm{d}\boldsymbol{\Omega}'\, \mathcal{W}_t[\boldsymbol{\Omega}, \boldsymbol{\Omega}'] P_t(\boldsymbol{\Omega}')$$

where

$$\mathcal{W}_t[\boldsymbol{\Omega}, \boldsymbol{\Omega}'] = N \big\langle \langle \delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}(\boldsymbol{J} + \Delta \boldsymbol{J})] \rangle_{\xi} - \delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}(\boldsymbol{J})] \big\rangle_{\Omega', t}$$

with the so-called sub-shell (or conditional) average $\langle \cdots \rangle_{\Omega', t}$, defined as

$$\langle f(\boldsymbol{J}) \rangle_{\Omega, t} = \frac{\int \mathrm{d}\boldsymbol{J}\, p_t(\boldsymbol{J}) \delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}(\boldsymbol{J})] f(\boldsymbol{J})}{\int \mathrm{d}\boldsymbol{J}\, p_t(\boldsymbol{J}) \delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}(\boldsymbol{J})]}.$$

It is possible to make various assumptions regarding the scaling behaviour of our observables at time $t = 0$, but once this has been specified the scaling at subsequent times is determined by the dynamics. We make the natural assumption that $Q(0) = \mathcal{O}(1)$ so that, in accordance with our assumptions regarding the statistics of $\boldsymbol{B}$, we have $R(0) = \mathcal{O}(N^{-1/2})$. We suppose that $S(0) = \mathcal{O}(N^{1/2})$, the maximum permitted by the Schwarz inequality.

In this context it is worth remarking that in the idealized case of zero bias, Hebbian learning is known to outperform the perceptron learning rule; but in the more realistic situation of even moderately biased data the Hebbian rule fails miserably. For example, if we assume that $S(0)$ is $\mathcal{O}(1)$, and that Q, R are initially $\mathcal{O}(1)$, it follows from the learning rule (or from the methods which we apply below to the perceptron learning rule) that in the initial evolution of the Hebbian system $\mathrm{d}S/\mathrm{d}\tau = \eta a^2 K(\lambda_2/\sigma)$, where $\tau = Nt$, so that S rapidly diverges and no learning takes place; the student vector $\boldsymbol{J}$ cannot break away from its alignment to the bias vector. We shall show, however, that the perceptron has no problem coping with extreme initial conditions such as $S(0) = \mathcal{O}(N^{1/2})$, and that in due course effective learning occurs. The Hebbian example also serves to show that, even if we were to choose the weaker initial scaling $S(0) = \mathcal{O}(N^0)$, dependent on the specific choice we make for the learning rule, the order parameter S might well be driven towards $S = \mathcal{O}(N^{1/2})$ states.

A systematic exploration of the possible scaling scenarios reveals the following†. For the perceptron learning rule and for the initial scaling conditions as specified above, the only self-consistent solution of the macroscopic equations is one describing a situation where the system passes through three phases $\{\mathrm{I}, \mathrm{II}, \mathrm{III}\}$ defined by time-scales $t = \{\tau N^{-1/2}, \tau N^{-1}, \tau\}$, in which our observables are $\mathcal{O}(1)$ quantities in all three phases, with the exception of S which is $\mathcal{O}(N^{1/2})$ in phase I. We will write $S = \widetilde{S} N^{1/2}$ in phase I, with $\widetilde{S} = \mathcal{O}(N^0)$, and formulate our phase I equations in terms of $\widetilde{S}$ rather than S. The number of iterations m is related to the original time t by $m = Nt$ so that the number of iterations up to time τ, in each of the three phases, is given by $m = \{\tau N^{1/2}, \tau N^0, \tau N\}$. We incorporate these scaling properties into our equations in each of the three phases, by working henceforth only with $\mathcal{O}(N^0)$ time units τ and $\mathcal{O}(N^0)$ observables $\boldsymbol{\Omega}$, which satisfy

$$\frac{\mathrm{d}}{\mathrm{d}\tau} P_\tau(\boldsymbol{\Omega}) = \int \mathrm{d}\boldsymbol{\Omega}'\, \mathcal{W}_\tau[\boldsymbol{\Omega}, \boldsymbol{\Omega}'] P_\tau(\boldsymbol{\Omega}') \tag{7}$$

† For brevity, in this paper we will only describe the resulting self-consistent solution, which is indeed perfectly consistent with the observations in numerical simulations such as in figure 1.

with

$$\mathcal{W}_\tau[\boldsymbol{\Omega},\boldsymbol{\Omega}'] = F_{\mathrm{I,II,III}}\,\langle\,\langle\delta[\boldsymbol{\Omega}-\boldsymbol{\Omega}(\boldsymbol{J}+\Delta\boldsymbol{J})]\rangle_\xi - \delta[\boldsymbol{\Omega}-\boldsymbol{\Omega}(\boldsymbol{J})]\,\rangle_{\boldsymbol{\Omega}',t}$$
$$= \frac{F_{\mathrm{I,II,III}}}{(2\pi)^3}\left\langle \int \mathrm{d}\hat{\boldsymbol{\Omega}}\,\mathrm{e}^{\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}}\left\{\langle \mathrm{e}^{-\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}(\boldsymbol{J}+\Delta\boldsymbol{J})}\rangle_\xi - \mathrm{e}^{-\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}(\boldsymbol{J})}\right\}\right\rangle_{\boldsymbol{\Omega}',t} \tag{8}$$

and $F_{\mathrm{I}} = N^{1/2}$, $F_{\mathrm{II}} = N^0 \quad F_{\mathrm{III}} = N$. In a subsequent stage it will be convenient to write $\Delta\boldsymbol{J} = \boldsymbol{k} + \boldsymbol{k}'$, where

$$\boldsymbol{k} = \frac{\eta}{2N}\boldsymbol{A}[\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}) - \mathrm{sgn}(\boldsymbol{J}\cdot\boldsymbol{\xi})] \qquad \boldsymbol{k}' = \frac{\eta}{2N}\boldsymbol{v}[\mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}) - \mathrm{sgn}(\boldsymbol{J}\cdot\boldsymbol{\xi})] \tag{9}$$

so that

$$\Delta\boldsymbol{J}\cdot\boldsymbol{A} = \tfrac{1}{2}\eta a^2[\mathrm{sgn}(\lambda_2+y) - \mathrm{sgn}(\lambda_1+x)] + \eta a\frac{z}{2\sqrt{N}}[\mathrm{sgn}(\lambda_2+y) - \mathrm{sgn}(\lambda_1+x)]. \tag{10}$$

We are now in a position to discuss the dynamics in each of the three phases in which different scaling laws apply.

4. Phase I: elimination of bias-induced activation

In phase I we define the $\mathcal{O}(N^0)$ observables $\boldsymbol{\Omega} = (\widetilde{S}, Q, R) = (\boldsymbol{J}\cdot\boldsymbol{A}/\sqrt{N}, Q, R)$ and $F_{\mathrm{I}} = \sqrt{N}$. Upon expanding the exponential $\mathrm{e}^{-\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}(\boldsymbol{J}+\Delta\boldsymbol{J})}$ in powers of $\Delta\boldsymbol{J}$ we obtain from equation (8)

$$\mathcal{W}_\tau[\boldsymbol{\Omega},\boldsymbol{\Omega}'] = -\frac{1}{(2\pi)^3}\int \mathrm{d}\hat{\boldsymbol{\Omega}}\,\langle \mathrm{e}^{\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot[\boldsymbol{\Omega}-\boldsymbol{\Omega}(\boldsymbol{J})]}\{\cdots\}_{\mathrm{I}}\rangle_{\boldsymbol{\Omega}',\tau}$$

where

$$\{\cdots\}_{\mathrm{I}} = \mathrm{i}\sum_{i\mu}\left\langle N^{1/2}\Delta J_i\frac{\partial\Omega_\mu}{\partial J_i}\right\rangle_\xi \hat{\Omega}_\mu + \frac{1}{2}\mathrm{i}\sum_{ij\mu}\left\langle N^{1/2}\Delta J_i\Delta J_j\frac{\partial^2\Omega_\mu}{\partial J_i\partial J_j}\right\rangle_\xi \hat{\Omega}_\mu$$
$$+\frac{1}{2}\sum_{ij\mu\nu}\left\langle N^{1/2}\Delta J_i\Delta J_j\frac{\partial\Omega_\mu}{\partial J_i}\frac{\partial\Omega_\nu}{\partial J_j}\right\rangle_\xi \hat{\Omega}_\mu\hat{\Omega}_\nu + \mathcal{O}(N^{-1}).$$

A straightforward calculation using equation (9) and the two averages $\langle\mathrm{sgn}(\lambda_2+y)\rangle = K(\lambda_2/\sigma)$ and $\langle\mathrm{sgn}(\lambda_1+x)\rangle = K(\lambda_1/\sigma) = \mathrm{sgn}(\widetilde{S})$ (which is valid for large N in phase I) now gives

$$\{\cdots\}_{\mathrm{I}} = \frac{1}{2}\mathrm{i}\eta a^2\left[K\left(\frac{\lambda_2}{\sigma}\right) - \mathrm{sgn}(\widetilde{S})\right]\hat{\Omega}_1 + \mathrm{i}\eta\widetilde{S}\left[K\left(\frac{\lambda_2}{\sigma}\right) - \mathrm{sgn}(\widetilde{S})\right]\hat{\Omega}_2.$$

We can now apply equation (7) to compute the time derivative of the probability density $P_\tau(\boldsymbol{\Omega})$. Note that the sub-shell average $\langle\cdots\rangle_{\boldsymbol{\Omega}',\tau}$ involves an integration over all $\boldsymbol{J}$ for which $\boldsymbol{\Omega}(\boldsymbol{J}) = \boldsymbol{\Omega}'$ (in a distributional sense) so in calculating the relevant integrals we may effectively replace $\boldsymbol{\Omega}(\boldsymbol{J})$ by $\boldsymbol{\Omega}'$ at appropriate stages. For example, $\int \mathrm{d}\hat{\boldsymbol{\Omega}}\,\hat{\Omega}_j\mathrm{e}^{\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot[\boldsymbol{\Omega}-\boldsymbol{\Omega}(\boldsymbol{J})]} = \mathrm{i}(2\pi)^3\partial_{j'}\delta[\boldsymbol{\Omega}-\boldsymbol{\Omega}']$, where $\partial_{j'}$ denotes differentiation with respect to Ω'_j. We now find for $P_\tau(\widetilde{S}, Q, R)$ a Liouville equation

$$\frac{\mathrm{d}}{\mathrm{d}\tau}P_\tau(\widetilde{S}, Q, R) = -\frac{\partial}{\partial\widetilde{S}}\left[\frac{\eta a^2}{2}\left[K\left(\frac{\lambda_2}{\sigma}\right) - \mathrm{sgn}(\widetilde{S})\right]P_\tau(\widetilde{S}, Q, R)\right]$$
$$-\frac{\partial}{\partial Q}\left[\eta\widetilde{S}\left[K\left(\frac{\lambda_2}{\sigma}\right) - \mathrm{sgn}(\widetilde{S})\right]P_\tau(\widetilde{S}, Q, R)\right]$$

with the deterministic solution $P_\tau(\widetilde{S}, Q, R) = \delta[\widetilde{S}-\widetilde{S}(\tau)]\,\delta[Q-Q(\tau)]\,\delta[R-R(\tau)]$, where the actual deterministic trajectory $\{\widetilde{S}(\tau), Q(\tau), R(\tau)\}$ is the solution of the coupled flow equations

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\widetilde{S} = \frac{1}{2}\eta a^2\left[K\left(\frac{\lambda_2}{\sigma}\right) - \mathrm{sgn}(\widetilde{S})\right] \qquad \frac{\mathrm{d}}{\mathrm{d}\tau}Q = \eta\widetilde{S}\left[K\left(\frac{\lambda_2}{\sigma}\right) - \mathrm{sgn}(\widetilde{S})\right] \qquad \frac{\mathrm{d}}{\mathrm{d}\tau}R = 0.$$

It follows that $\widetilde{S}(\tau) = \widetilde{S}(0) + \frac{1}{2}\eta a^2\tau[K(\lambda_2/\sigma) - \mathrm{sgn}(\widetilde{S})]$. We see that $\widetilde{S}$ is driven to zero in times $\tau = \tau_\pm$ (with $\pm$ referring to the cases $\widetilde{S}_0 > 0$ and $\widetilde{S}_0 < 0$, respectively), which are given by

$$\tau_\pm = \frac{2|\widetilde{S}_0|}{\eta a^2(1 \mp K(\lambda_2/\sigma))}.$$

Irrespective of the value of $\widetilde{S}_0$, the system seeks to eliminate any strong alignment of the learning vector $\boldsymbol{J}$ relative to the bias vector $\boldsymbol{A}$. This is clearly confirmed by numerical simulations. Our equation for Q also readily integrates to give $Q = Q_0 + [\widetilde{S}^2 - \widetilde{S}_0^2]/a^2$. We see that the length $J = \sqrt{Q}$ of the student weight vector decreases and that $J \to [J_0^2 - \widetilde{S}_0^2/a^2]^{1/2}$ as $\tau \to \tau_\pm$. Again, this is confirmed by numerical simulations. The equation $\mathrm{d}R/\mathrm{d}\tau = 0$ implies that ωJ is constant in phase I. As can be clearly seen in figure 1, no learning takes place in this phase, since expression (6) for E_g reduces to $E_g = \frac{1}{2}[1 - \mathrm{sgn}(S)K(\lambda_2/\sigma)]$ in the limit $|\lambda_1| \to \infty$ (note $\lambda_1 = S/J$). However, at times τ approaching $\tau_\pm$ it is no longer valid to argue that S is $\mathcal{O}(\sqrt{N})$; it is now $\mathcal{O}(N^0)$ and we enter the scaling regime of phase II.

5. Phase II: transition to error correction

As shown in the previous section, S is an $\mathcal{O}(N^0)$ quantity in phase II, and it is also clear that $\{Q, R\}$ are $\mathcal{O}(N^0)$ at the start of phase II. In phase II (and, as we will see, also in phase III) we have to consider the observables $\boldsymbol{\Omega} = (S, \boldsymbol{\Phi})$, with $\boldsymbol{\Phi} = (Q, R)$; the reason for this slight departure from our phase I terminology will soon become clear. We can now express equation (8) as

$$\mathcal{W}_\tau[\boldsymbol{\Omega}, \boldsymbol{\Omega}'] = F_{\mathrm{II},\mathrm{III}} \int \frac{\mathrm{d}\hat{\boldsymbol{\Omega}}}{(2\pi)^3}\, \mathrm{e}^{\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}} \left\langle \langle \mathrm{e}^{-\mathrm{i}\hat{S}S(\boldsymbol{J}+\Delta\boldsymbol{J}) - \mathrm{i}\sum_{\mu=1}^{2}\hat{\Phi}_\mu\Phi_\mu(\boldsymbol{J}+\Delta\boldsymbol{J})}\rangle_\xi - \mathrm{e}^{-\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}(\boldsymbol{J})}\right\rangle_{\boldsymbol{\Omega}',\tau}.$$

Here

$$\Phi_\mu(\boldsymbol{J}+\Delta\boldsymbol{J}) = \Phi_\mu(\boldsymbol{J}) + \sum_i \Delta J_i \frac{\partial\Phi_\mu}{\partial J_i} + \frac{1}{2}\sum_{ij}\Delta J_i\Delta J_j\frac{\partial^2\Phi_\mu}{\partial J_i\partial J_j}$$

(this expansion is exact, since $\{Q, R\}$ are quadratic and linear functions, respectively). Substituting and expanding the exponential gives

$$\mathcal{W}_\tau[\boldsymbol{\Omega}, \boldsymbol{\Omega}'] = \left\langle F_{\mathrm{II},\mathrm{III}} \int \frac{\mathrm{d}\hat{\boldsymbol{\Omega}}}{(2\pi)^3}\, \mathrm{e}^{\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot[\boldsymbol{\Omega}-\boldsymbol{\Omega}(\boldsymbol{J})]}\langle \mathrm{e}^{-\mathrm{i}\hat{S}\Delta\boldsymbol{J}\cdot\boldsymbol{A}} - 1\rangle_\xi - \int \frac{\mathrm{d}\hat{\boldsymbol{\Omega}}}{(2\pi)^3}\, \mathrm{e}^{\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot[\boldsymbol{\Omega}-\boldsymbol{\Omega}(\boldsymbol{J})]}\{\cdots\}_{\mathrm{II},\mathrm{III}}\right\rangle_{\boldsymbol{\Omega}'}$$

where

$$\{\cdots\}_{\mathrm{II},\mathrm{III}} = \mathrm{i}F_{\mathrm{II},\mathrm{III}}\sum_{i\mu}\left\langle \Delta J_i\frac{\partial\Phi_\mu}{\partial J_i}\mathrm{e}^{-\mathrm{i}\hat{S}\Delta\boldsymbol{J}\cdot\boldsymbol{A}}\right\rangle_\xi \hat{\Phi}_\mu + \frac{1}{2}\mathrm{i}F_{\mathrm{II},\mathrm{III}}\sum_{ij\mu}\left\langle \Delta J_i\Delta J_j\frac{\partial^2\Phi_\mu}{\partial J_i\partial J_j}\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\right\rangle_\xi \hat{\Phi}_\mu$$
$$+\frac{1}{2}F_{\mathrm{II},\mathrm{III}}\sum_{ij\mu\nu}\left\langle \Delta J_i\Delta J_j\frac{\partial\Phi_\mu}{\partial J_i}\frac{\partial\Phi_\nu}{\partial J_j}\mathrm{e}^{-\mathrm{i}\hat{S}\Delta\boldsymbol{J}\cdot\boldsymbol{A}}\right\rangle_\xi \hat{\Phi}_\mu\hat{\Phi}_\nu + \mathcal{O}(N^{-3/2}). \qquad (11)$$

Note that whereas it is valid to expand $e^{-i\hat{\Phi}\cdot\Phi(J+\Delta J)}$ in the manner just described, we cannot treat $e^{-i\hat{S}S(J+\Delta J)}$ in the same way since $\sum_i \Delta J_i(\partial S/\partial J_i) = \boldsymbol{A}\cdot\Delta\boldsymbol{J} = \mathcal{O}(N^0)$ in phases II and III. Equations (7) and (11) form the basis for our study of phases II and III.

The time-scale τ in phase II is related to t via $t = \tau N^{-1}$, so that $F_{\rm II} = N^0$, but although this phase is of short duration it has an important role as regards the stochastic evolution of the bias overlap parameter S. It is straightforward to show that the third term in equation (11) makes no contribution in the limit of large N. Moreover, in the very short phase II we may approximate $\Delta\boldsymbol{J}\cdot\boldsymbol{A}$ by $\boldsymbol{k}\cdot\boldsymbol{A}$ (10). Referring to the details and notation in the appendix we have

$$\langle e^{-i\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\rangle = 1 - E_g + e^{i\eta a^2\hat{S}} I_1(-\lambda_1, \lambda_2, -\omega) + e^{-i\eta a^2\hat{S}} I_1(\lambda_1, -\lambda_2, -\omega) \qquad (12)$$

and we then find that in phase II

$$\mathcal{W}_\tau[\boldsymbol{\Omega}, \boldsymbol{\Omega}'] = I_1(-\lambda_1, \lambda_2, -\omega)\delta[S - S' + \eta a^2]\delta[\boldsymbol{\Phi} - \boldsymbol{\Phi}'] + I_1(\lambda_1, -\lambda_2, \omega)\delta[S - S' - \eta a^2]\delta[\boldsymbol{\Phi} - \boldsymbol{\Phi}'] - E_g\delta[\boldsymbol{\Omega} - \boldsymbol{\Omega}'].$$

With substitution into (7) and repetition of the arguments used for phase I we find that Q and R remain constant in phase II, whilst the conditional distribution $P_\tau(S|Q, R)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\tau}P_\tau(S|Q, R) = I_1(\lambda_1(S^-), -\lambda_2, -\omega)P_\tau(S^-|Q, R) + I_1(-\lambda_1(S^+), \lambda_2, -\omega)P_\tau(S^+|Q, R) - E_g(S, Q, R)P_\tau(S|Q, R)$$

where $S^\pm = S \pm \eta a^2$. The distribution equilibrates, on the relevant time-scale, to a stationary distribution $P(S|Q, R)$ given as the solution of

$$E_g(S, Q, R)P(S|Q, R) = I_1(\lambda_1(S^-), -\lambda_2, -\omega)P(S^-|Q, R) + I_1(-\lambda_1(S^+), \lambda_2, -\omega)P(S^+|Q, R).$$

Using relation (5) we find that this equilibrium condition can be written as $A(S) + B(S) = A(S^+) + B(S^-)$, where $A(S) = I_1(-\lambda_1(S), \lambda_2, -\omega)P(S|Q, R)$ and $B(S) = I_1(\lambda_1(S), -\lambda_2, -\omega)P(S|Q, R)$. One can easily show by taking Fourier transforms that it is satisfied by $B(S) = A(S^+)$, the correctness of which is evident by substitution. In this phase the permissible values of S are those which differ from some initial value $S(0)$ by an integral multiple of ηa^2. Upon writing the allowed values of S as $S_n = S(0) + n\eta a^2$, we immediately obtain $P(S|Q, R) = \sum_{n=-\infty}^{\infty} w(S_{n+1}|Q, R)\,\delta[S - S_n]$, where

$$w(S_{n+1}|Q, R) = \frac{I_1(\lambda_1(n), -\lambda_2, -\omega)}{I_1(-\lambda_1(n+1), \lambda_2, -\omega)}\, w(S_n|Q, R) \qquad (13)$$

with I_1 as given in (A1). Equation (13) fully determines the quasi-stationary distribution $P(S|Q, R)$. Comparison with numerical simulations shows very satisfactory agreement (see, e.g., figure 2). The above picture is also in line with our intuition, since in a single step the change in S is given by

$$\Delta S = \Delta\boldsymbol{J}\cdot\boldsymbol{A} = \tfrac{1}{2}\eta a^2[\mathrm{sgn}(\lambda_2 + y) - \mathrm{sgn}(\lambda_1 + x)] + \tfrac{1}{2}\eta a\frac{z}{\sqrt{N}}[\mathrm{sgn}(\lambda_2 + y) - \mathrm{sgn}(\lambda_1 + x)].$$

Provided we can neglect the $N^{-1/2}$ term in this expression, which is true on the time-scale of phase II, we see that in a single update $\Delta S \in \{0, \pm\eta a^2\}$. However, if the $N^{-1/2}$ term could be neglected indefinitely this would imply that, far into the future, the system would retain a memory of its initial conditions. In fact, the term $\frac{1}{2}\eta a z[\mathrm{sgn}(\lambda_2 + y) - \mathrm{sgn}(\lambda_1 + x)]/\sqrt{N}$ represents a random walk superposed on the quasi-stationary distribution found for S in phase II.

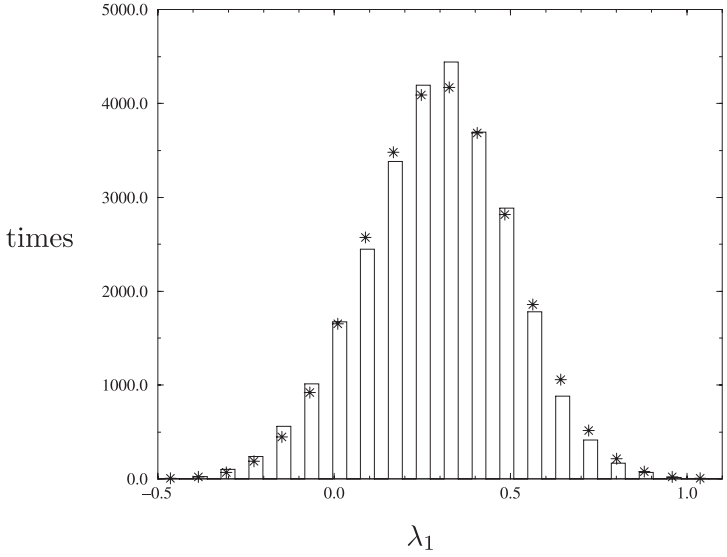

Figure 2. Histogram of values for $\lambda_1 = S/J$ as measured in a single simulation experiment with $N = 400\,000$, $\eta = 1$ and $a = \frac{1}{2}$, during the interval $t \in [0, 0.07]$. Here phase I is absent, by virtue of the choice $S(0) = 0$, and $\lambda_2 = 0.287$. The stars indicate the predicted occurrence probabilities as calculated from (13). Over a short observation time-scale the observed distribution for λ_1 is truly discrete: no values of λ_1 were found in between the centres of the histogram bars.

6. Phase III: error correction

As we enter phase III, where $F_{\rm III} = N$, the above 'random walk' term will come to have a significant role after approximately N iterations†, leading to a modified probability distribution which contains a diffusion term: $S_n \to S_n + s(t)$. The walk is given by

$$s(t) = \frac{\eta a}{2\sqrt{N}} \sum_{\mu=1}^{Nt} z(\mu)[\mathrm{sgn}(\lambda_2 + y(\mu)) - \mathrm{sgn}(\lambda_1(\mu) + x(\mu))]$$

in an obvious notation, where the fields $z(\mu)$ are, as we have seen earlier, independent of (x, y). The random walk addition $s(t)$ has a mean of zero, and a variance given by

$$\langle s^2(t)\rangle = \frac{\eta^2 a^2 \sigma^2}{2N} \sum_{\mu=1}^{Nt} \langle [1 - \mathrm{sgn}(\lambda_2 + y(\mu))\,\mathrm{sgn}(\lambda_1(\mu) + x(\mu))]\rangle = t(\eta a \sigma)^2 \langle E_g\rangle \tag{14}$$

where $\langle E_g\rangle$ is to be interpreted as a time average of E_g over phase III, up to time t.

In order to extract the macroscopic laws in phase III we will now have to analyse this diffusion effect carefully, starting from equation (11). The details of this analysis are given in appendix B, where we show that for large N the macroscopic distribution in phase III will again be of the form $P_\tau(S, Q, R) = P_\tau(S|Q, R)\delta[Q - Q(\tau)]\delta[R - R(\tau)]$, but now with the deterministic values $\{Q(\tau), R(\tau)\}$ given as the solution of the coupled equations

$$\frac{\mathrm{d}}{\mathrm{d}\tau} Q = \eta\sqrt{Q} \int \mathrm{d}S\,(K_1 + L_1 + M_1) P_\tau(S|Q, R) + \tfrac{1}{2}\eta^2 \int \mathrm{d}S\,(K_3 + L_3 + M_3) P_\tau(S|Q, R) \tag{15}$$

† We are grateful to Peter Sollich for pointing this out.

$$\frac{\mathrm{d}}{\mathrm{d}\tau}R = \tfrac{1}{2}\eta \int \mathrm{d}S\,(K_2 + L_2 + M_2)P_\tau(S|Q,R). \tag{16}$$

The factors $\{K_i, L_i, M_i\}$, defined in appendix B, are indeed functions of S (via λ_1) and of $\{Q, R\}$. The origin and meaning of these two equations can be appreciated more clearly by writing them in the following, somewhat more appealing, form (without as yet specifying the learning rule $\mathcal{F}[J, u, T]$):

$$\frac{\mathrm{d}}{\mathrm{d}\tau}Q = 2\eta J \int \mathrm{d}S\,P_\tau(S|Q,R)\,\langle(\lambda_1 + x)\,\mathrm{sgn}(\lambda_2 + y)\mathcal{F}[\sqrt{Q}, \lambda_1 + x, \mathrm{sgn}(\lambda_2 + y)]\rangle$$
$$+\eta^2 \int \mathrm{d}S\,P_\tau(S|Q,R)\,\langle\mathcal{F}^2[\sqrt{Q}, \lambda_1 + x, \mathrm{sgn}(\lambda_2 + y)]\rangle$$

$$\frac{\mathrm{d}}{\mathrm{d}\tau}R = \eta \int \mathrm{d}S\,P_\tau(S|Q,R)\,\langle|y|\mathcal{F}[\sqrt{Q}, \lambda_1 + x, \mathrm{sgn}(\lambda_2 + y)]\rangle$$

(see [13] for details). Although equations (15) and (16) are superficially similar to the equations which we derived in phase I, we now have a situation in which functions of S are weighted with respect to the probability distribution $P_\tau(S|Q,R)$ which satisfies a partial differential equation derived from equation (B4) (in appendix B) by integration over Q and R, namely

$$\frac{\mathrm{d}}{\mathrm{d}\tau}P_\tau(S|Q,R) = N\big[I_1(-\lambda_1(S^+,\lambda_2,-\omega)P_\tau(S^+|Q,R) + I_1(\lambda_1(S^-,-\lambda_2,-\omega)P_\tau(S^-|Q,R)$$
$$-E_g(S,Q,R)P_\tau(S|Q,R)\big]$$
$$+\frac{1}{2}\eta^2a^2\sigma^2\bigg[\frac{\partial^2}{\partial S^2}[I_1(-\lambda_1(S^+),\lambda_2,-\omega)P_\tau(S^+|Q,R)]$$
$$+\frac{\partial^2}{\partial S^2}[I_1(\lambda_1(S^-),-\lambda_2,-\omega)P_\tau(S^-|Q,R)]\bigg]. \tag{17}$$

Equations (15)–(17), together with the definitions of the shorthand version $\{K_i, L_i, M_i\}$ as given in appendix B, provide an exact and closed set of equations for the macroscopic dynamics in phase III, in terms of the observables $\{S, Q, R\}$. In the large-N limit, Q and R satisfy deterministic equations, as in conventional no-bias theories, but S remains stochastic throughout phase III. Furthermore, the persistent appearance of the factor λ_2 (which depends on the actual realization of the teacher weights) induces sample-to-sample fluctuations. An example of the result of solving the coupled equations (15)–(17) numerically (via a numerical realization, i.e. Monte Carlo, of the conditional stochastic process (17) for S) is shown in figure 3, and compared with numerical simulations of the underlying microscopic perceptron learning process. The agreement between theory and experiment is quite satisfactory.

7. Asymptotics of the generalization error

A full numerical study of our equations (15)–(17) would be difficult, but these equations undergo a great simplification, permitting further analysis, if we make the approximation $P_\tau(S|Q,R) = \delta[S - \langle S\rangle]$, and assume that $\lambda_1(\langle S\rangle) = \lambda_2$; numerical simulations confirm the validity of the replacement of λ_1 by λ_2 *on average* in phase III. In this approximation equations (15) and (16) become

$$\frac{\mathrm{d}}{\mathrm{d}\tau}Q = \eta\sqrt{Q}[K_1 + L_1 + M_1] + \tfrac{1}{2}\eta^2[K_3 + L_3 + M_3] \qquad \frac{\mathrm{d}}{\mathrm{d}\tau}R = \tfrac{1}{2}[K_2 + L_2 + M_2].$$

Note that $K_1 + L_1 + M_1 = \lambda_1[(A_1 + B_1 + C_1) - (A_2 + B_2 + C_2)] + (A_3 + B_3 + C_3) - (A_4 + B_4 + C_4)$. Referring to appendix A for the relevant expressions for $\{A_i, B_i, C_i\}$

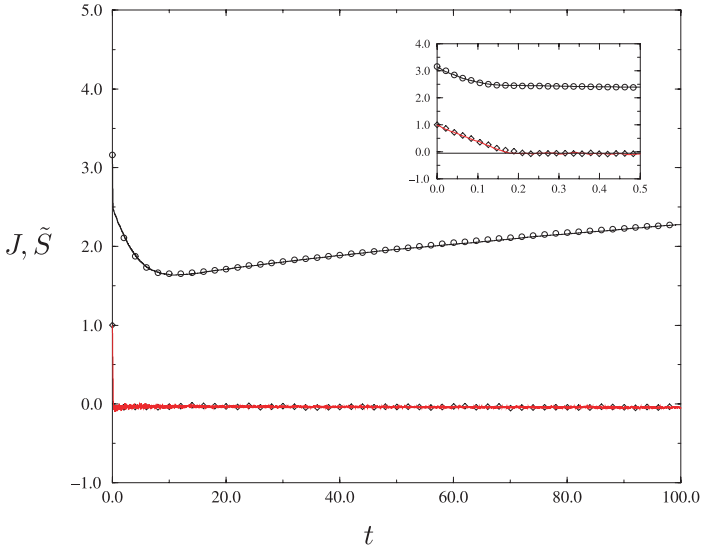

Figure 3. Evolution of the order parameters $J = |\boldsymbol{J}|$ (upper) and $\tilde{S} = S/\sqrt{N} = \boldsymbol{J}\cdot\boldsymbol{A}/\sqrt{N}$ (lower), for $a = \frac{1}{2}$ and $\eta = 1$. Markers indicate simulation results (for $N = 1000$), the full curve is the theoretical curve obtained by numerical solution of equations (15)–(17). Initial conditions: $Q(0) = 10$, $R(0) = 0$ and $\tilde{S}(0) = 1$. On the time-scale $t = \mu/N$ only phase III is visible. The inset shows a magnified view of the initial stage of the process, where phase I can be observed.

in terms of the integrals $I_1(\lambda_1, \lambda_2, \omega)$ and $I_2(\lambda_1, \lambda_2, \omega)$, and using the identity $K(\alpha) = \int_{-\infty}^{\infty} Dy\, K((\alpha - \omega y)/\sqrt{1-\omega^2})$, we find that in the approximation $\lambda_1 = \lambda_2$ the following identities hold:

$$A_1 + B_1 + C_1 = K(\lambda_2/\sigma) \qquad A_2 + B_2 + C_2 = K(\lambda_2/\sigma)$$

$$A_3 + B_3 + C_3 = \sqrt{\frac{2}{\pi}}\omega\sigma \exp\left[-\frac{\lambda_2^2}{2\sigma^2}\right] \qquad A_4 + B_4 + C_4 = \sqrt{\frac{2}{\pi}}\sigma \exp\left[-\frac{\lambda_2^2}{2\sigma^2}\right]$$

$$K_1 + L_1 + M_1 = -\sqrt{\frac{2}{\pi}}\sigma(1-\omega)\exp\left[-\frac{\lambda_2^2}{2\sigma^2}\right].$$

$$K_2 + L_2 + M_2 = (A_5 + B_5 + C_5) - (A_6 + B_6 + C_6) = \sqrt{\frac{2}{\pi}}\sigma(1-\omega)\exp\left[-\frac{\lambda_2^2}{2\sigma^2}\right]$$

$$K_3 + L_3 + M_3 = 1 - \int_{-\lambda_2/\sigma}^{\infty} Dy\, K\left(\frac{\lambda_2 + \omega\sigma y}{\sigma\sqrt{1-\omega^2}}\right) + \int_{\lambda_2/\sigma}^{\infty} Dy\, K\left(\frac{\lambda_2 - \omega\sigma y}{\sigma\sqrt{1-\omega^2}}\right) = 2E_g$$

and equations (15) and (16) therefore become (upon rewriting the equation for Q in terms of $J = \sqrt{Q}$)

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau}J &= -\frac{\eta}{\sqrt{2\pi}}\sigma(1-\omega)\exp\left[-\frac{\lambda_2^2}{2\sigma^2}\right] + \frac{\eta^2}{2J}E_g \\ \frac{\mathrm{d}}{\mathrm{d}\tau}R &= \frac{\eta}{\sqrt{2\pi}}\sigma(1-\omega)\exp\left[-\frac{\lambda_2^2}{2\sigma^2}\right]. \end{aligned} \tag{18}$$

The corresponding equation for $\omega = R/J$ is

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\omega = \frac{\eta}{J\sqrt{2\pi}}\sigma(1-\omega^2)\exp\left[-\frac{\lambda_2^2}{2\sigma^2}\right] - \frac{\omega\eta^2}{2J}E_g \tag{19}$$

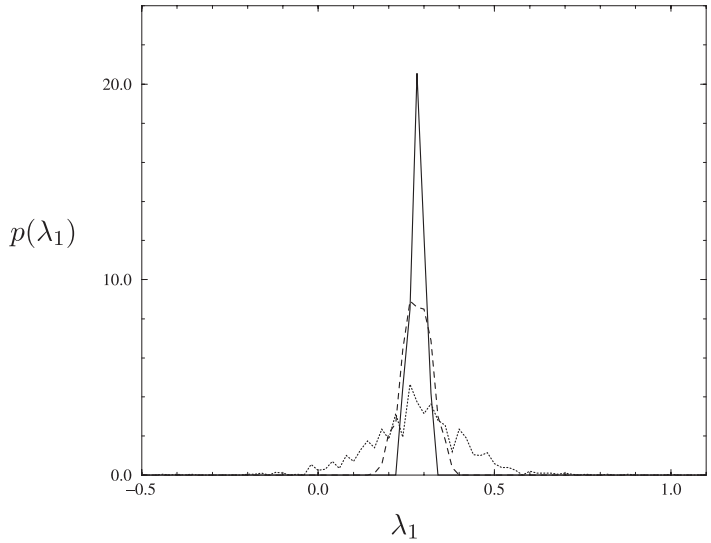

Figure 4. Distribution $p(\lambda_1)$ of $\lambda_1 = S/J$, as measured during a single simulation run, with $N = 1000$, $a = \frac{1}{2}$ and $\eta = 1$, over the time intervals [0, 100] (dotted curve), [950, 1050] (broken curve) and [9900, 10 000] (full curve). One observes that the fluctuations in λ_1 are reduced to zero, as time progresses.

which is to be solved in combination with (6). The numerical solution of these equations is found to be in very good agreement with the results of numerical simulations, even for finite times; however, it is relevant to consider what basis exists for making the approximation $\lambda_1 = \lambda_2$, other than the fact that it works. We have already observed that the probability distribution for S in phase III is a random walk superposed on the underlying discrete distribution which emerged in phase II. Equation (14) indicates that the random walk, reflected in the diffusion terms in equation (17), could, in principle, lead to a large variance for S, were this random walk is not coupled to the underlying discrete distribution via equation (17). The discrete distribution and the random walk, however, are found to interact in such a way that the fluctuations actually tend to zero in the limit $\tau \to \infty$; this is confirmed by the results of numerical simulations which show that the fluctuations in $\lambda_1 = S/J$ decrease with time and that on average λ_1 tends to λ_2 (see figure 4). In a single step the *average change* in S is equal to

$$\tfrac{1}{2}\eta a^2\langle[\mathrm{sgn}(\lambda_2 + y) - \mathrm{sgn}(\lambda_1 + x)]\rangle + \tfrac{1}{2}\eta a\frac{\langle z\rangle}{\sqrt{N}}\langle[\mathrm{sgn}(\lambda_2 + y) - \mathrm{sgn}(\lambda_1 + x)]\rangle$$

$$= \tfrac{1}{2}\eta a^2\left[K\left(\frac{\lambda_2}{\sigma}\right) - K\left(\frac{\lambda_1}{\sigma}\right)\right]$$

so, as the fluctuations in S diminish, we do indeed expect that λ_1 will tend to λ_2.

We will now use the coupled equations (6) and (19) to derive an asymptotic expression for the generalization error E_g. Differentiation of (6) with respect to ω gives

$$\frac{\partial E_g}{\partial \omega} = -\frac{\mathrm{e}^{-\frac{1}{2}\beta^2}}{\pi\sqrt{1-\omega^2}}\mathrm{e}^{-\frac{1}{2}\beta^2(1-\omega)/(1+\omega)}$$

with the constant $\beta = \lambda_2/\sigma$. Changing the variable to $\omega = \cos\theta$, and expanding for $\theta \to 0$ gives

$$E_g = \pi^{-1}\mathrm{e}^{-\frac{1}{2}\beta^2}\int_0^\theta \mathrm{d}u\,\mathrm{e}^{-\frac{1}{2}\beta^2\tan^2(u/2)} = \pi^{-1}\mathrm{e}^{-\frac{1}{2}\beta^2}\left[\theta - \frac{\beta^2\theta^3}{24} + \mathcal{O}(\theta^5)\right]. \quad (20)$$

Equation (18) for J and equation (19) for ω can now be written

$$\frac{\mathrm{d}J}{\mathrm{d}\tau} = -\frac{\eta\sigma}{\sqrt{2\pi}}(1-\cos\theta)\mathrm{e}^{-\frac{1}{2}\beta^2} + \frac{\eta^2 E_g}{2J}$$

$$-J\sin\theta\frac{\mathrm{d}\theta}{\mathrm{d}\tau} = \frac{\eta\sigma}{\sqrt{2\pi}}\sin^2\theta\,\mathrm{e}^{-\frac{1}{2}\beta^2} - \frac{\eta^2 E_g\cos\theta}{2J}.$$

Using the expansion $\tan\theta = \theta + \frac{1}{3}\theta^3 + \mathcal{O}(\theta^5)$ we then expand our previous equations for the evolution of J and θ, giving

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\theta = \mathrm{e}^{-\frac{1}{2}\beta^2}\left\{-\frac{\eta\sigma\theta}{J\sqrt{2\pi}} + \frac{\eta^2}{2\pi J^2} - \frac{\eta^2\theta^2\rho}{2\pi J^2}\right\} + \mathcal{O}(\theta^4) \qquad \text{with} \quad \rho = \tfrac{1}{24}\beta^2 + \tfrac{1}{3}$$

$$\frac{\mathrm{d}}{\mathrm{d}\tau}J = \mathrm{e}^{-\frac{1}{2}\beta^2}\left\{-\frac{\eta\sigma\theta^2}{2\sqrt{2\pi}} + \frac{\eta^2}{2\pi J}\left[\theta - \tfrac{1}{24}\beta^2\theta^3\right]\right\} + \mathcal{O}(\theta^5).$$

Upon making the asymptotic ansatz $J = A/\theta$, the equation for J can now be expressed so as to give a second equation for θ. The two resulting equations for $\mathrm{d}\theta/\mathrm{d}\tau$ are

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\theta = \frac{\eta\sigma\theta^4}{2A\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\beta^2} - \frac{\eta^2\theta^4}{2A^2\pi}\left[1 - \frac{\beta^2\theta^2}{24} + \mathcal{O}(\theta^4)\right]\mathrm{e}^{-\frac{1}{2}\beta^2}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\theta = -\frac{\eta\sigma\theta^2}{A\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\beta^2} + \frac{\eta^2\theta^2}{2\pi A^2}\mathrm{e}^{-\frac{1}{2}\beta^2} + \mathcal{O}(\theta^4).$$

Consistency requires that A be given by $A = \eta/\sigma\sqrt{2\pi}$. The asymptotic equation for θ subsequently becomes $\mathrm{d}\theta/\mathrm{d}\tau = -\frac{1}{2}\sigma^2\theta^4\mathrm{e}^{-\frac{1}{2}\beta^2}$, from which we obtain the asymptotic power law $\theta = k\tau^\alpha$, where $\alpha = -\frac{1}{3}$ and $k^3 = 2\mathrm{e}^{\frac{1}{2}\beta^2}/3\sigma^2$. Combining this, finally, with (20) we then obtain, recalling that in phase III one simply has $\tau = m/N = t$:

$$E_g(t) = \rho(a)\mathrm{e}^{-\lambda_2^2/3\sigma^2}t^{-1/3} \quad (t\to\infty) \qquad \rho(a) = \left[\frac{2}{3\pi^3}\right]^{1/3}(1-a^2)^{-1/3}. \quad (21)$$

Note that the power of τ occurring in this expression is the same as the power which appears in the asymptotic form of the generalization error in the conventional no-bias theory; the coefficient is, however, different, but reduces to the familiar form in the case of zero bias, where $a = \lambda_2 = 0$ and $\sigma = 1$. Moreover, our prediction of the asymptotic form of E_g is in excellent agreement with the results of numerical simulations. This is evident from figure 5, where we show the observed function $\rho(a)$, defined as $\rho(a) = \lim_{t\to\infty} E_g(t)t^{1/3}\mathrm{e}^{\lambda_2^2/3\sigma^2}$, versus the theoretical prediction as given in (21). Note that the dependence of (21) on the teacher-bias overlap $\lambda_2 = \boldsymbol{B}\cdot\boldsymbol{A}$ implies sample-to-sample fluctuations.

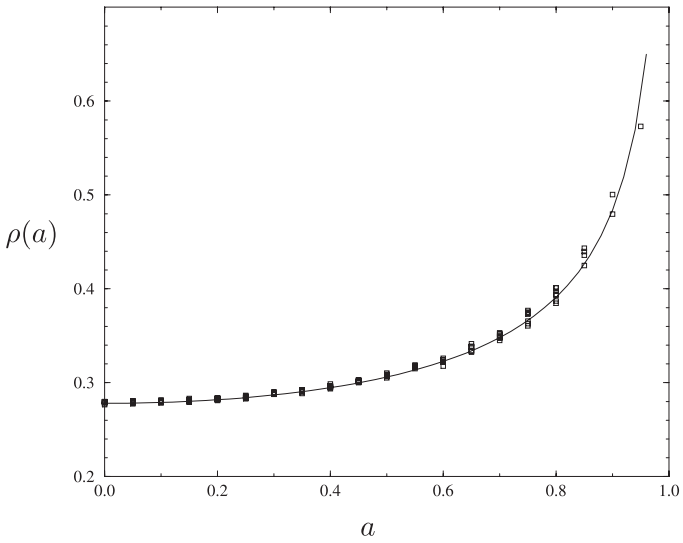

Figure 5. Comparison of $\rho(a)$ as found in simulations ($N = 1000$ and $\eta = 1$, see the main text for details of its definition), for various values of the teacher-bias overlap $\lambda_2 = \boldsymbol{B} \cdot \boldsymbol{A}$ (squares), with the theoretical prediction (21) (full curve).

8. Discussion

We have studied analytically the dynamics of on-line learning in nonlinear perceptrons, trained according to the perceptron rule, for the scenario of having structurally biased, i.e. $\mathcal{O}(N^0)$, input data. The bias changes the learning process qualitatively, inducing three distinct phases (with different scaling properties) and persistent stochastic as well as sample-to-sample fluctuations in the generalization error, even for $N \to \infty$. At a theoretical level, the need to introduce an extra order parameter S (the projection of the student weight vector in the direction of the bias) which is neither deterministic nor self-averaging, makes the analysis considerably more involved than that of the idealized bias-free case. In the third and final phase, in which most of the learning takes place, we have obtained a set of exact closed equations which involve the conditional probability density of S. However, because of their complicated nature, an exact analytic solution of these equations appears to be out of the question, as is also generally the case in the more familiar no-bias scenarios. Nevertheless, we have found that an approximate (and much simpler) version of our equations yields results which are in excellent agreement with numerical simulations. We show that the asymptotic power law for the generalization error is largely preserved, with the bias showing up only in the pre-factor. At various stages throughout out calculations we have compared the predictions of our macroscopic dynamic equations with the results of numerical simulations of the underlying (microscopic) learning process, which consistently showed excellent agreement.

Although in this paper we have confined ourselves to the perceptron learning rule, it is clear that our analysis is in no way restricted to this particular rule, and can be applied to other rules such as the AdaTron learning rule, where $\Delta \boldsymbol{J} = \frac{\eta}{2N} \boldsymbol{\xi} [\mathrm{sgn}(\boldsymbol{B} \cdot \boldsymbol{\xi}) - \mathrm{sgn}(\boldsymbol{J} \cdot \boldsymbol{\xi})] |\boldsymbol{J} \cdot \boldsymbol{\xi}|$; one could even study optimal learning rates and optimal learning rules, generalizing [14] to the case of having $a \neq 0$. Preliminary studies of the AdaTron learning rule with structurally biased data show, for instance, that the simple result (13), describing the phase II distribution

in the case of the perceptron, is replaced by the integral equation

$$E_g P_\tau(\widetilde{S}|Q,R) = \eta a^2 J \int_{-\lambda_1}^{\infty} \mathrm{d}\rho\, G(\rho,\lambda_2)\, P_\tau(\widetilde{S} + (\lambda_1+\rho)\eta a^2 J)|Q,R)$$
$$+\eta a^2 J \int_{\lambda_1}^{\infty} \mathrm{d}\rho\, G(\rho,-\lambda_2) P_\tau(\widetilde{S} + (\lambda_1-\rho)\eta a^2 J)|Q,R)$$

where G is defined by

$$G(x,\lambda_2) = \frac{\mathrm{e}^{-(x^2/2\sigma^2)}}{2\sigma\sqrt{2\pi}}\left[1 - K\left(\frac{\lambda_2+\omega x}{\sigma\sqrt{1-\omega^2}}\right)\right].$$

The discrete distribution which in the present paper we found for the perceptron in phase II no longer applies in the AdaTron case, and is replaced by a continuous distribution which satisfies the above integral equation. The analysis of the AdaTron in the case of biased data is more complicated than for the perceptron, as might have been expected from the nature of the AdaTron learning rule, but much of the work which we have presented for the perceptron can be carried through and the results will be published later. There is also scope for a more detailed mathematical investigation of the partial differential equation which we derived to describe the conditional probability distribution $P_\tau(S|Q,R)$ for the perceptron, but this is likely to be difficult, and is beyond the scope of the present paper.

Acknowledgments

It is a pleasure to thank Peter Sollich and Michael Biehl for helpful comments and discussions. JAFH wishes to thank King's College London Association for financial support.

Appendix A. Integrals and averages

We recall that the function K is defined by $K(x) = \mathrm{erf}(x/\sqrt{2})$. In terms of this definition note that $\int_\tau^\infty \mathrm{d}\zeta\, \mathrm{e}^{-\frac{1}{2}\zeta^2} = \frac{1}{2}\sqrt{2\pi}[1-K(\tau)]$. We now proceed to list various integrals which occur in our calculations, or are referred to in the text, and where appropriate outline a brief derivation. Recall that the joint distribution of $(x,y) = (\hat{\boldsymbol{J}}\cdot\boldsymbol{v}, \boldsymbol{B}\cdot\boldsymbol{v})$ is given by

$$p(x,y) = [2\pi\sigma^2\sqrt{1-\omega^2}]^{-1} \exp\left[-\frac{1}{2}\frac{[x^2 - 2\omega xy + y^2]}{\sigma^2(1-\omega^2)}\right]$$

where $\sigma^2 = 1 - a^2$ and $\omega = \boldsymbol{B}\cdot\hat{\boldsymbol{J}}$. We then find that

$$\begin{aligned} I_1(\lambda_1,\lambda_2,\omega) &= \int_{\lambda_1}^{\infty} \mathrm{d}x \int_{\lambda_2}^{\infty} \mathrm{d}y\, p(x,y) \\ &= \int_{\lambda_2}^{\infty} \frac{\mathrm{d}y}{2\pi\sigma} \exp\left[-\frac{y^2}{2\sigma^2}\right] \int_{(\lambda_1-\omega y)/(\sigma\sqrt{1-\omega^2})}^{\infty} \mathrm{d}\zeta\, \mathrm{e}^{-\frac{1}{2}\zeta^2} \\ &= \frac{1}{4}\left[1 - K\left(\frac{\lambda_2}{\sigma}\right)\right] - \frac{1}{2}\int_{\lambda_2/\sigma}^{\infty} Dy\, K\left(\frac{\lambda_1 - \omega\sigma y}{\sigma\sqrt{1-\omega^2}}\right) \end{aligned} \tag{A1}$$

and similarly

$$I_2(\lambda_1,\lambda_2,\omega)=\int_{\lambda_1}^{\infty}\mathrm{d}x\,x\int_{\lambda_2}^{\infty}\mathrm{d}y\,p(x,y)=\int_{\lambda_1}^{\infty}\frac{\mathrm{d}x\,x}{2\sigma\sqrt{2\pi}}\exp\left[-\frac{x^2}{2\sigma^2}\right]$$
$$-\int_{\lambda_1}^{\infty}\frac{\mathrm{d}x\,x}{2\sqrt{2\pi}\sigma}\exp\left[-\frac{x^2}{2\sigma^2}\right]K\left(\frac{\lambda_2-\omega x}{\sigma\sqrt{1-\omega^2}}\right)$$
$$=\frac{\sigma}{2\sqrt{2\pi}}\exp\left[-\frac{\lambda_1^2}{2\sigma^2}\right]\left[1-K\left(\frac{\lambda_2-\omega\lambda_1}{\sigma\sqrt{1-\omega^2}}\right)\right]$$
$$+\frac{\omega\sigma}{2\sqrt{2\pi}}\exp\left[-\frac{\lambda_2^2}{2\sigma^2}\right]\left[1-K\left(\frac{\lambda_1-\omega\lambda_2}{\sigma\sqrt{1-\omega^2}}\right)\right].$$

The following averages with respect to the distribution $p(x,y)$ are easily calculated:

$$\langle\mathrm{sgn}(\lambda_1+x)\rangle=K\left(\frac{\lambda_1}{\sigma}\right)\qquad\langle x\,\mathrm{sgn}(\lambda_2+y)\rangle=\sqrt{\frac{2}{\pi}}\omega\sigma\exp\left[-\frac{\lambda_2^2}{2\sigma^2}\right]$$

$$\langle x\,\mathrm{sgn}(\lambda_1+x)\rangle=\sqrt{\frac{2}{\pi}}\sigma\exp\left[-\frac{\lambda_1^2}{2\sigma^2}\right]$$

$$\langle\mathrm{sgn}(\lambda_1+x)\,\mathrm{sgn}(\lambda_2+y)\rangle=I_1(\lambda_1,\lambda_2,\omega)-I_1(-\lambda_1,\lambda_2,-\omega)-I_1(\lambda_1,-\lambda_2,-\omega)$$
$$+I_1(-\lambda_1,-\lambda_2,\omega)$$
$$=\int_{-\lambda_2/\sigma}^{\infty}Dy\,K\left(\frac{\lambda_1+\omega\sigma y}{\sigma\sqrt{1-\omega^2}}\right)-\int_{\lambda_2/\sigma}^{\infty}Dy\,K\left(\frac{\lambda_1-\omega\sigma y}{\sigma\sqrt{1-\omega^2}}\right).$$

Finally, in studying phases II and III we require the following averages:

$$\langle\mathrm{sgn}(\lambda_2+y)\,\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\rangle=A_1+B_1\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2}+C_1\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}$$
$$\langle\mathrm{sgn}(\lambda_1+x)\,\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\rangle=A_2+B_2\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2}+C_2\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}$$
$$\langle x\,\mathrm{sgn}(\lambda_2+y)\,\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\rangle=A_3+B_3\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2}+C_3\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}$$
$$\langle x\,\mathrm{sgn}(\lambda_1+x)\,\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\rangle=A_4+B_4\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2}+C_4\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}$$
$$\langle y\,\mathrm{sgn}(\lambda_2+y)\,\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\rangle=A_5+B_5\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2}+C_5\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}$$
$$\langle y\,\mathrm{sgn}(\lambda_1+x)\,\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\rangle=A_6+B_6\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2}+C_6\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}$$
$$\langle\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\rangle=A_7+B_7\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2}+C_7\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}$$
$$\langle\mathrm{sgn}(\lambda_1+x)\,\mathrm{sgn}(\lambda_2+y)\,\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\rangle=A_8+B_8\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2}+C_8\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}$$

where

$$A_1=-[I_1(\lambda_1,\lambda_2,\omega)-I_1(-\lambda_1,-\lambda_2,\omega)]$$
$$B_1=-I_1(-\lambda_1,\lambda_2-\omega)$$
$$C_1=I_1(\lambda_1,-\lambda_2,-\omega)$$
$$A_2=-[I_1(\lambda_1,\lambda_2,\omega)-I_1(-\lambda_1,-\lambda_2,\omega)]$$
$$B_2=I_1(-\lambda_1,\lambda_2,-\omega)$$
$$C_2=-I_1(\lambda_1,-\lambda_2,-\omega)$$

$$
\begin{aligned}
A_3 &= [I_2(\lambda_1, \lambda_2, \omega) + I_2(-\lambda_1, -\lambda_2, \omega)] \\
B_3 &= -I_2(-\lambda_1, \lambda_2, -\omega) \\
C_3 &= -I_2(\lambda_1, -\lambda_2, -\omega) \\
A_4 &= [I_2(\lambda_1, \lambda_2, \omega) + I_2(-\lambda_1, -\lambda_2, \omega)] \\
B_4 &= -I_2(-\lambda_1, \lambda_2, -\omega) \\
C_4 &= I_2(\lambda_1, -\lambda_2, -\omega) \\
A_5 &= [I_2(\lambda_2, \lambda_1, \omega) + I_2(-\lambda_2, -\lambda_1, \omega)] \\
B_5 &= I_2(\lambda_2, -\lambda_1, -\omega) \\
C_5 &= I_2(-\lambda_2, \lambda_1, -\omega) \\
A_6 &= [I_2(\lambda_2, \lambda_1, \omega) + I_2(-\lambda_2, -\lambda_1, \omega)] \\
B_6 &= -I_2(\lambda_2, -\lambda_1, -\omega) \\
C_6 &= -I_2(-\lambda_2, \lambda_1, -\omega) \\
A_7 &= 1 - E_g \\
B_7 &= I_1(-\lambda_1, \lambda_2, -\omega) \\
C_7 &= I_1(\lambda_1, -\lambda_2, -\omega) \\
A_8 &= [I_1(\lambda_1, \lambda_2, \omega) + I_1(-\lambda_1, -\lambda_2, \omega)] \\
B_8 &= -I_1(-\lambda_1, \lambda_2, -\omega) \\
C_8 &= -I_1(\lambda_1, -\lambda_2, -\omega).
\end{aligned}
$$

All these formulae may be established by elementary methods. For example,

$$
\begin{aligned}
\langle x\,\mathrm{sgn}(\lambda_2 + y)\,\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\rangle &= \int \mathrm{d}x\,\mathrm{d}y\,x\,p(x, y)\,\mathrm{sgn}(\lambda_2 + y)\mathrm{e}^{-\frac{1}{2}\mathrm{i}\hat{S}\eta a^2[\mathrm{sgn}(\lambda_2+y)-\mathrm{sgn}(\lambda_1+x)]} \\
&= \left[\int_{-\infty}^{-\lambda_1} \mathrm{d}x\,x + \int_{-\lambda_1}^{\infty} \mathrm{d}x\,x\right]\left[\int_{-\lambda_2}^{\infty} \mathrm{d}y\,\mathrm{e}^{-\frac{1}{2}\mathrm{i}\hat{S}\eta a^2(1-\mathrm{sgn}(\lambda_1+x))}\,p(x, y)\right] \\
&\quad -\left[\int_{-\infty}^{-\lambda_1} \mathrm{d}x\,x + \int_{-\lambda_1}^{\infty} \mathrm{d}x\,x\right]\left[\int_{\lambda_2}^{\infty} \mathrm{d}y\,\mathrm{e}^{\frac{1}{2}\mathrm{i}\hat{S}\eta a^2(1+\mathrm{sgn}(\lambda_1+x))}\,p(x, -y)\right] \\
&= \int_{\lambda_1}^{\infty} \mathrm{d}x\,x \int_{\lambda_2}^{\infty} \mathrm{d}y\,p(x, y) - \int_{-\lambda_1}^{\infty} \mathrm{d}x\,x \int_{\lambda_2}^{\infty} \mathrm{d}y\,\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2}\,p(x, -y) \\
&\quad - \int_{\lambda_1}^{\infty} \mathrm{d}x\,x \int_{-\lambda_2}^{\infty} \mathrm{d}y\,\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}\,p(x, -y) + \int_{-\lambda_1}^{\infty} \mathrm{d}x\,x \int_{-\lambda_2}^{\infty} \mathrm{d}y\,p(x, y) \\
&= I_2(\lambda_1, \lambda_2, \omega) - \mathrm{e}^{\mathrm{i}\hat{S}\eta a^2} I_2(-\lambda_1, \lambda_2, -\omega) - \mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2} I_2(\lambda_1, -\lambda_2, -\omega) \\
&\quad + I_2(-\lambda_1, -\lambda_2, \omega) \\
&= A_3 + B_3\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2} + C_3\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}
\end{aligned}
$$

as required.

Appendix B. Analysis of macroscopic distribution in phase III

Here we give the details of our analysis of the macroscopic distribution $P_\tau(S, Q, R)$ in phase III, starting from equation (11). We note that, in phase III:

$$\mathrm{e}^{-\mathrm{i}\hat{S}\Delta \boldsymbol{J}\cdot\boldsymbol{A}} = \mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\left\{1 - \frac{\mathrm{i}\eta a z\hat{S}}{2\sqrt{N}}[\mathrm{sgn}(\lambda_2 + y) - \mathrm{sgn}(\lambda_1 + x)]\right.$$
$$\left. - \frac{(\eta a z\hat{S})^2}{4N}[1 - \mathrm{sgn}(\lambda_1 + x)\,\mathrm{sgn}(\lambda_2 + y)] + \cdots\right\}.$$

The terms which we neglected are $\mathcal{O}(N^{-2})$, since when performing averages over the training set the average of the z^3 term is zero. Equation (11) now yields

$$\mathcal{W}_\tau[\boldsymbol{\Omega}, \boldsymbol{\Omega}'] = N\int \frac{\mathrm{d}\hat{\boldsymbol{\Omega}}}{(2\pi)^3}\left\langle \mathrm{e}^{\mathrm{i}\hat{\Omega}\cdot[\Omega - \Omega(J)]}\left\langle \mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}} - 1\right.\right.$$
$$\left. - \frac{(\eta a\sigma\hat{S})^2}{4N}\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}[1 - \mathrm{sgn}(\lambda_1 + x)\,\mathrm{sgn}(\lambda_2 + y)]\right\rangle_\xi$$
$$\left. - \int \frac{\mathrm{d}\hat{\boldsymbol{\Omega}}}{(2\pi)^3}\,\mathrm{e}^{\mathrm{i}\hat{\Omega}\cdot[\Omega - \Omega(J)]}\{\cdots\}_{\mathrm{III}}\right\rangle_{\Omega'} \tag{B1}$$

where

$$\{\cdots\}_{\mathrm{III}} = \mathrm{i}N\sum_{i\mu}\left\langle k_i\frac{\partial\Phi_\mu}{\partial J_i}\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\right\rangle_\xi\hat{\Phi}_\mu + \frac{1}{2}\mathrm{i}N\sum_{ij\mu}\left\langle k_ik_j\frac{\partial^2\Phi_\mu}{\partial J_i\partial J_j}\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\right\rangle_\xi\hat{\Phi}_\mu$$
$$+\frac{1}{2}N\sum_{ij\mu\nu}\left\langle k_ik_j\frac{\partial\Phi_\mu}{\partial J_i}\frac{\partial\Phi_\nu}{\partial J_j}\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\right\rangle_\xi\hat{\Phi}_\mu\hat{\Phi}_\nu. \tag{B2}$$

We showed in appendix A that

$$\int \mathrm{d}x\,\mathrm{d}y\;p(x, y)\,\mathrm{sgn}(\lambda_1 + x)\,\mathrm{sgn}(\lambda_2 + y)\,\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}} = I_1(\lambda_1, \lambda_2, \omega) - I_1(-\lambda_1, \lambda_2, -\omega)\,\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2}$$
$$-I_1(\lambda_1, -\lambda_2, -\omega)\,\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2} + I_1(-\lambda_1, -\lambda_2, \omega)$$

so that

$$\int \mathrm{d}x\,\mathrm{d}y\;p(x, y)\,\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}[1 - \mathrm{sgn}(\lambda_1 + x)\,\mathrm{sgn}(\lambda_2 + y)]$$
$$= 2\left[\mathrm{e}^{\mathrm{i}\hat{S}\eta a^2}I_1(-\lambda_1, \lambda_2, -\omega) + \mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}I_1(\lambda_1, -\lambda_2, -\omega)\right]$$

by virtue of equation (12) and the fact that $I_1(\lambda_1, \lambda_2, \omega) + I_1(-\lambda_1, -\lambda_2, \omega) = 1 - E_g$. Bearing in mind the sub-shell average we may write

$$\int \mathrm{d}\hat{S}\,\hat{S}^2\,\mathrm{e}^{\mathrm{i}\hat{S}[S \pm \eta a^2 - S(J)]} = -2\pi\frac{\partial^2}{\partial S'^2}\delta[S \pm \eta a^2 - S'].$$

Upon combining equations (7), (12) and (B1) we find that in phase III the joint probability density $P_\tau(S, Q, R)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\tau}P_\tau(S, Q, R) = N\left[I_1(-\lambda_1(S^+), \lambda_2, -\omega)P_\tau(S^+, Q, R)\right.$$
$$\left. + I_1(\lambda_1(S^-), -\lambda_2, -\omega)P_\tau(S^-, Q, R) - E_g(S, Q, R)P_\tau(S, Q, R)\right]$$

$$+\frac{1}{2}\eta^2a^2\sigma^2\int \mathrm{d}\boldsymbol{\Omega}'\,P_\tau(\boldsymbol{\Omega}')\Big[I_1(-\lambda_1(S'),\lambda_2,-\omega)\frac{\partial^2}{\partial S'^2}\delta[S+\eta a^2-S']$$

$$+I_1(\lambda_1(S'),-\lambda_2,-\omega)\frac{\partial^2}{\partial S'^2}\delta[S-\eta a^2-S']\Big]\delta[\boldsymbol{\Phi}-\boldsymbol{\Phi}']$$

$$-\int\frac{\mathrm{d}\boldsymbol{\Omega}'}{(2\pi)^3}\,P_\tau(\boldsymbol{\Omega}')\left\langle\int \mathrm{d}\hat{\boldsymbol{\Omega}}\,\mathrm{e}^{\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot[\boldsymbol{\Omega}-\boldsymbol{\Omega}(J)]}\{\cdots\}_{\mathrm{III}}\right\rangle_{\boldsymbol{\Omega}'}$$

where $\{\cdots\}_{\mathrm{III}}$ is given by equation (B2), and hence

$$\frac{\mathrm{d}}{\mathrm{d}\tau}P_\tau(S,Q,R)=N\big[I_1(-\lambda_1(S^+),\lambda_2,-\omega)P_\tau(S^+,Q,R)$$

$$+I_1(\lambda_1(S^-),-\lambda_2,-\omega)P_\tau(S^-,Q,R)-E_g(S,Q,R)P_\tau(S,Q,R)\big]$$

$$+\frac{1}{2}\eta^2a^2\sigma^2\Big[\frac{\partial^2}{\partial S^2}[I_1(-\lambda_1(S^+),\lambda_2,-\omega)P_\tau(S^+,Q,R)]$$

$$+\frac{\partial^2}{\partial S^2}[I_1(\lambda_1(S^-),-\lambda_2,-\omega)P_\tau(S^-,Q,R)]\Big]$$

$$-\int\frac{\mathrm{d}\boldsymbol{\Omega}'}{(2\pi)^3}\,P_\tau(\boldsymbol{\Omega}')\int \mathrm{d}\hat{\boldsymbol{\Omega}}\big\langle\mathrm{e}^{\mathrm{i}\hat{\boldsymbol{\Omega}}\cdot[\boldsymbol{\Omega}-\boldsymbol{\Omega}(J)]}\{\cdots\}_{\mathrm{III}}\big\rangle_{\boldsymbol{\Omega}'}. \tag{B3}$$

As regards the evaluation of $\{\cdots\}_{\mathrm{III}}$ we note that

$$N\sum_{i\mu}\left\langle k_i\frac{\partial\Phi_\mu}{\partial J_i}\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\right\rangle_{\boldsymbol{\xi}}\hat{\Phi}_\mu=\eta J\big\langle(\lambda_1+x)[\mathrm{sgn}(\lambda_2+y)-\mathrm{sgn}(\lambda_1+x)]\,\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\big\rangle\hat{\Phi}_1$$

$$+\tfrac{1}{2}\eta\langle(\lambda_2+x)[\mathrm{sgn}(\lambda_2+y)-\mathrm{sgn}(\lambda_1+x)]\,\mathrm{e}^{-\mathrm{i}\hat{s}\boldsymbol{k}\cdot\boldsymbol{A}}\rangle\hat{\Phi}_2$$

$$=\eta J[K_1+L_1\mathrm{e}^{\mathrm{i}\hat{s}\eta a^2}+M_1\mathrm{e}^{-\mathrm{i}\hat{s}\eta a^2}]\hat{\Phi}_1+\tfrac{1}{2}\eta[K_2+L_2\mathrm{e}^{\mathrm{i}\hat{s}\eta a^2}+M_2\mathrm{e}^{-\mathrm{i}\hat{S}\eta a^2}]\hat{\Phi}_2$$

in which

$$K_1=\lambda_1(A_1-A_2)+(A_3-A_4)\qquad K_2=\lambda_2(A_1-A_2)+(A_5-A_6)$$

$$L_1=\lambda_1(B_1-B_2)+(B_3-B_4)\qquad L_2=\lambda_2(B_1-B_2)+(B_5-B_6)$$

$$M_1=\lambda_1(C_1-C_2)+(C_3-C_4)\qquad M_2=\lambda_2(C_1-C_2)+(C_5-C_6)$$

and A_i, B_i and C_i are functions defined in appendix A, and expressed in terms of the integrals $I_1(\lambda_1,\lambda_2,\omega)$ and $I_2(\lambda_1,\lambda_2,\omega)$. In a similar way we find that

$$N\sum_{ij\mu}\left\langle k_ik_j\frac{\partial^2\Phi_\mu}{\partial J_i\partial J_j}\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\right\rangle_{\boldsymbol{\xi}}\hat{\Phi}_\mu=\eta^2\big\langle[1-\mathrm{sgn}(\lambda_1+x)\,\mathrm{sgn}(\lambda_2+y)]\,\mathrm{e}^{\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\big\rangle\hat{\Phi}_1$$

$$=\eta^2\big[K_3+L_3\mathrm{e}^{\mathrm{i}\hat{s}\eta a^2}+M_3^{-\mathrm{i}\hat{s}\eta a^2}\big]\hat{\Phi}_1$$

where $K_3=A_7-A_8$, $L_3=B_7-B_8$ and $M_3=C_7-C_8$. Note that the term

$$N\sum_{ij\mu\nu}\left\langle k_ik_j\frac{\partial\Phi_\mu}{\partial J_i}\frac{\partial\Phi_\nu}{\partial J_j}\mathrm{e}^{-\mathrm{i}\hat{S}\boldsymbol{k}\cdot\boldsymbol{A}}\right\rangle_{\boldsymbol{\xi}}\hat{\Phi}_\mu\hat{\Phi}_\nu$$

makes no contribution to $\mathcal{W}_\tau[\boldsymbol{\Omega},\boldsymbol{\Omega}']$ in the limit of large N. Using equations (B2) and (B3) we can now carry out the remaining integrations using standard formulae from distribution

theory, as described for earlier phases, and find that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}P_\tau(S,Q,R) = N\big[&I_1(-\lambda_1(S^+),\lambda_2,-\omega)P_\tau(S^+,Q,R)\\ &+I_1(\lambda_1(S^-),-\lambda_2,-\omega)P_\tau(S^-,Q,R) - E_g(S,Q,R)P_\tau(S,Q,R)\big]\\ &+\tfrac{1}{2}\eta^2a^2\sigma^2\bigg[\frac{\partial^2}{\partial S^2}[I_1(-\lambda_1(S^+),\lambda_2,-\omega)P_\tau(S^+,Q,R)]\\ &+\frac{\partial^2}{\partial S^2}[I_1(\lambda_1(S^-),-\lambda_2,-\omega)P_\tau(S^-,Q,R)]\bigg]\\ &-\frac{\partial}{\partial Q}\big[\eta J[K_1P_\tau(S,Q,R)+L_1P_\tau(S^+,Q,R)+M_1P_\tau(S^-,Q,R)]\\ &+\tfrac{1}{2}\eta^2[K_3P_\tau(S,Q,R)+L_3P_\tau(S^+,Q,R)+M_3P_\tau(S^-,Q,R)]\big]\\ &-\frac{\partial}{\partial R}\big[\tfrac{1}{2}\eta[K_2P_\tau(S,Q,R)+L_2P_\tau(S^+,Q,R)+M_2P_\tau(S^-,Q,R)]\big]. \qquad \text{(B4)}\end{aligned}$$

Integration over S now gives, in combination with the relation $E_g = I_1(-\lambda_1,\lambda_2,-\omega) + I_1(\lambda_1,-\lambda_2,-\omega)$:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}P_\tau(Q,R) = &-\frac{\partial}{\partial Q}\bigg\{P_\tau(Q,R)\bigg[\eta J\int \mathrm{d}S\,(K_1+L_1+M_1)P_\tau(S|Q,R)\\ &+\frac{1}{2}\eta^2\int \mathrm{d}S\,(K_3+L_3+M_3)P_\tau(S|Q,R)\bigg]\bigg\}\\ &-\frac{\partial}{\partial R}\bigg\{P_\tau(Q,R)\bigg[\frac{1}{2}\eta\int \mathrm{d}S\,(K_2+L_2+M_2)P_\tau(S|Q,R)\bigg]\bigg\}\end{aligned}$$

which is a Liouville equation with solution $P_\tau(Q,R) = \delta[Q-Q(\tau)]\delta[R-R(\tau)]$, where the deterministic flow trajectories $(Q(\tau), R(\tau))$ are given as the solutions of (15) and (16), as claimed.

References

[1] Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan)
[2] Biehl M and Riegler P 1994 *Europhys. Lett.* **28** 525
[3] Mace C W H and Coolen A C C 1998 *Stat. Comput.* **8** 55
[4] Rae H C, Sollich P and Coolen A C C 1999 *Advances in Neural Information Processing Systems* vol 11, ed M S Kearns *et al* (Cambridge, MA: MIT Press)
[5] Rae H C, Sollich P and Coolen A C C 1999 *J. Phys. A: Math. Gen.* **32** 3321
[6] Mace C W H and Coolen A C C 2000 *Advances in Neural Information Processing Systems* vol 12, ed S A Solla, T K Leen and K R Müller (Cambridge, MA: MIT Press)
[7] Coolen A C C and Saad D 1998 *On-Line Learning in Neural Networks* ed D Saad (Cambridge: Cambridge University Press)
[8] Heimel J A F and Coolen A C C 2000 in preparation
[9] Li S and Wong K Y M 2000 *Advances in Neural Information Processing Systems* vol 12, ed S A Solla, T K Leen and K R Müller (Cambridge, MA: MIT Press)
[10] Marangi C, Solla S A, Biehl M and Riegler P 1995 *Proc. 7th Italian Workshop on Neural Nets (Vietri, 1995)* ed M Marinaro and R Tagliaferri (Singapore: World Scientific)
[11] Wendemuth A, Opper M and Kinzel W 1993 *J. Phys. A: Math. Gen.* **26** 3165
[12] Sollich P and Barber D 1998 *Neural Comput.* **10** 2201
[13] Heimel J A F, Rae H C and Coolen A C C 2000 *King's College London Preprint* KCL-MTH-00-26
[14] Kinouchi O and Caticha N 1992 *J. Phys. A: Math. Gen.* **25** 6243