Dynamics of Supervised Learning with Restricted Training Sets

A.C.C. Coolen † *and D. Saad* ‡

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

‡ Department of Computer Science and Applied Mathematics, Aston University
Aston Triangle, Birmingham B4 7ET, U.K.

Abstract

We study the dynamics of supervised learning in layered neural networks, in the regime where the size p of the training set is proportional to the number N of inputs. Here the local fields are no longer described by Gaussian distributions. We show how dynamical replica theory can be used to predict the evolution of macroscopic observables, including the relevant performance measures, incorporating the theory of complete training sets in the limit $p/N \to \infty$ as a special case. For simplicity we restrict ourselves here to single-layer networks and realizable tasks.

Contents

1 Introduction

In the last few years much progress has been made in the analysis of the dynamics of supervised learning in layered neural networks, using the strategy of statistical mechanics: by deriving from the microscopic dynamical equations a set of closed laws describing the evolution of suitably chosen macroscopic observables (dynamic order parameters) in the limit of an infinite system size [eg. Kinzel and Rujan (1990), Kinouchi and Caticha (1992), Biehl and Schwarze (1992,1995), Saad and Solla (1995)]. A recent review and more extensive guide to the relevant references can be found in Mace and Coolen (1998a). The main successful procedure developed so far is built on the following cornerstones:

- *The task to be learned is defined by a (possibly noisy) 'teacher', which is itself a layered neural network.* This induces a canonical set of dynamical order parameters, typically the (rescaled) overlaps between the various student weight vectors and the corresponding teacher weight vectors.
- *The number of network inputs is (eventually) taken to be infinitely large.* This ensures that fluctuations in mean-field observables will vanish and creates the possibility of using the central limit theorem.
- *The number of 'hidden' neurons is finite.* This prevents the number of order parameters from being infinite, and ensures that the cumulative impact of their fluctuations is insignificant.
- *The size of the training set is much larger than the number of updates made.* Each example presented is now different from those that have already been seen, such that the local fields will have Gaussian probability distributions, which leads to closure of the dynamic equations.

These are not ingredients to simplify the calculations, but vital conditions, without which the standard method fails. Although the assumption of an infinite system size has been shown not to be too critical (Barber et al, 1996), the other assumptions do place serious restrictions on the degree of realism of the scenarios that can be analyzed, and have thereby, to some extent, prevented the theoretical results from being used by practitioners.

In this paper we study the dynamics of learning in layered neural networks with restricted training sets, where the number p of examples ('questions' with corresponding 'answers') scales linearly with the number N of inputs, i.e. $p = \alpha N$. Here individual questions will re-appear during the learning process as soon as the number of weight updates made is of the order of the size of the training set. In the traditional models, where the duration of an update is defined as N^{-1}, this happens as soon as $t = \mathcal{O}(\alpha)$. At that point correlations develop between the weights and the questions in the training set, and the dynamics is of a spin-glass type, with the composition of the training set playing the role of 'quenched disorder'. The main consequence of this is that the central limit theorem no longer applies to the student's local fields, which are now described by non-Gaussian distributions. To demonstrate this we

trained (on-line) a perceptron with weights J_i on noiseless examples generated by a teacher perceptron with weights B_i, using the Hebb and AdaTron rules. We plotted in Fig. 1 the student and teacher fields, $x = \boldsymbol{J}\cdot\boldsymbol{\xi}$ and $y = \boldsymbol{B}\cdot\boldsymbol{\xi}$ respectively, where $\boldsymbol{\xi}$ is the input vector, for $p = N/2$ examples and at time $t = 50$. The marginal distribution $P(x)$ for $p = N/4$, at times $t = 10$ for the Hebb rule and $t = 20$ for the Adatron rule, is shown in Fig. 2. The non-Gaussian student field distributions observed in Figs. 1 and 2 induce a deviation between the training- and generalization errors, which measure the network performance on training and test examples, respectively. The former involves averages over the non-Gaussian field distribution, whereas the latter (which is calculated over *all* possible examples) still involves Gaussian fields.

The appearance of non-Gaussian fields leads to a breakdown of the standard formalism, based on deriving closed equations for a finite number of observables: the field distributions can no longer be characterized by a few moments, and the macroscopic laws must now be averaged over realizations of the training set. One could still try to use Gaussian distributions as large α approximations, see e.g. Sollich and Barber (1998), but it will be clear from Figs. 1 and 2 that a systematic theory will have to give up Gaussian distributions entirely. The first rigorous study of the dynamics of learning with restricted training sets in non-linear networks, via the calculation of generating functionals, was carried out by Horner (1992) for perceptrons with binary weights. In this paper we show how the formalism of dynamical replica theory (see e.g. Coolen et al, 1996) can be used successfully to predict the evolution of macroscopic observables for finite α, incorporating the infinite training set formalism as a special case, for $\alpha \to \infty$. Central to our approach is the derivation of a diffusion equation for the joint distribution of the student and teacher fields, which will be found to have Gaussian solutions only for $\alpha \to \infty$. For simplicity and transparency we restrict ourselves to single-layer systems and noise-free teachers. Application and generalization of our methods to multi-layer systems (Saad and Coolen, 1998) and learning scenarios involving 'noisy' teachers (Mace and Coolen, 1998b) are presently under way.

This presentation of preliminary results is organized as follows. In section 2 we derive a general Fokker-Planck equation describing the evolution of mean-field observables for $N \to \infty$. This allows us to identify the conditions for the latter to be described by closed deterministic laws. In section 3 we choose as our observables the field distribution $P[x, y]$, in addition to (the traditional) Q and R, and show that this set obeys deterministic laws. In order to close these laws we use the tools of dynamical replica theory. Details of the replica calculation are given in section 4, to be skipped by those primarily interested in results. In section 5 we show how in the limit $\alpha \to \infty$ (infinite training sets) the equations of the conventional theory are recovered. We finally work out our equations explicitly for the example of Hebbian learning with restricted training sets, and compare our predictions with exact results (derived from the microscopic equations by Rae et al, 1998) and with numerical simulations.

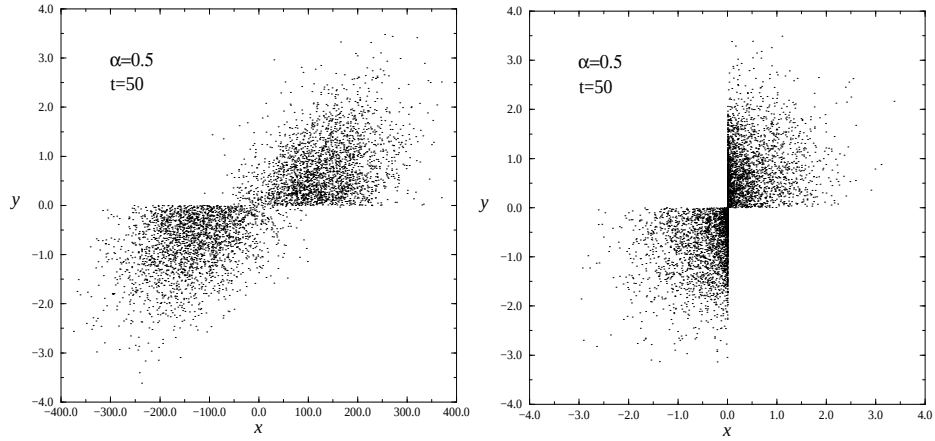

Fig. 1: Student and teacher fields (x, y) as observed during numerical simulations of on-line learning (learning rate $\eta = 1$) in a perceptron of size $N = 10,000$ at $t = 50$, using 'questions' from a restricted training set of size $p = \frac{1}{2}N$. Left: Hebbian learning. Right: AdaTron learning. Note: in the case of Gaussian field distributions one would have found spherically shaped plots.

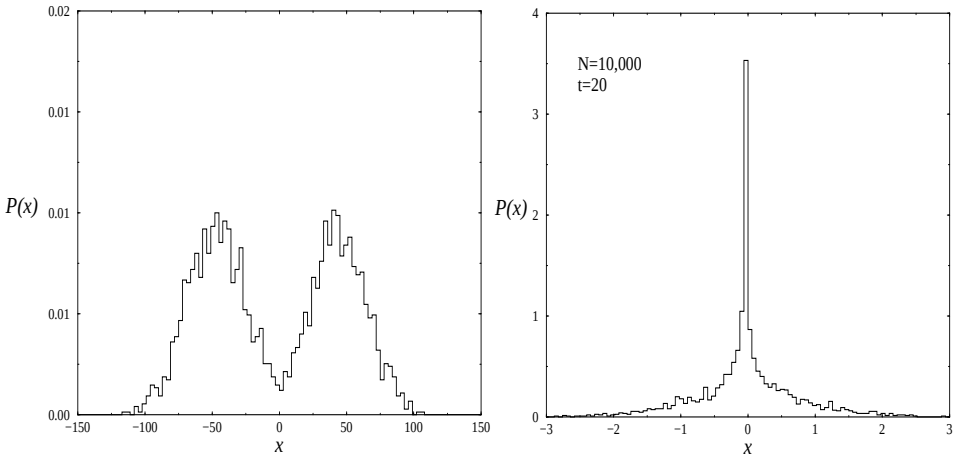

Fig. 2: Distribution $P(x)$ of student fields as observed during numerical simulations of on-line learning (learning rate $\eta = 1$) in a perceptron of size $N = 10,000$, using 'questions' from a restricted training set of size $p = \frac{1}{4}N$. Left: Hebbian learning, measured at $t = 10$. Right: AdaTron learning, measured at $t = 20$. Note: not only are these distributions distinctively non-Gaussian, they also appear to vary widely in their basic characteristics, depending on the learning rule used.

2 From Microscopic to Macroscopic Laws

2.1 Definitions

A student perceptron operates the following rule, which is parametrised by the weight vector $\boldsymbol{J} \in \Re^N$:

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

It tries to emulate the operation of a teacher perceptron, via an iterative procedure for updating its parameters $\boldsymbol{J}$. The teacher perceptron operates a similar rule, characterized by a given (fixed) weight vector $\boldsymbol{B} \in \Re^N$:

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

In order to do so, the student perceptron modifies its weight vector $\boldsymbol{J}$ accord-

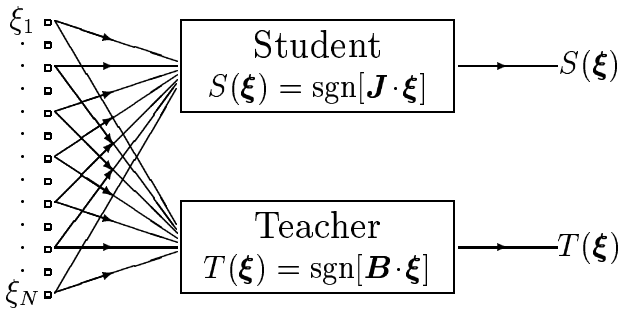

Fig. 3: Supervised learning in perceptrons.

ing to an iterative procedure, using examples of input vectors (or 'questions') $\boldsymbol{\xi}$, drawn at random from a fixed training set $\tilde{D} \subseteq D = \{-1,1\}^N$, and the corresponding values of the teacher outputs $T(\boldsymbol{\xi})$, see Fig. 3.

We consider the case where the training set is a randomly composed subset $\tilde{D} \subset D$, of size $|\tilde{D}| = p = \alpha N$ with $\alpha > 0$:

$$\tilde{D} = \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p\} \qquad\qquad p = \alpha N$$

We will denote averages over the training set $\tilde{D}$ and averages over the full question set D in the following way:

$$\langle \Phi(\boldsymbol{\xi}) \rangle_{\tilde{D}} = \frac{1}{|\tilde{D}|} \sum_{\boldsymbol{\xi}\in\tilde{D}} \Phi(\boldsymbol{\xi}) \qquad \text{and} \qquad \langle \Phi(\boldsymbol{\xi}) \rangle_{D} = \frac{1}{|D|} \sum_{\boldsymbol{\xi}\in D} \Phi(\boldsymbol{\xi}) \ .$$

We will analyze the following two classes of learning rules:

$$\begin{aligned} &\mathrm{on-line}: \quad \boldsymbol{J}(m+1) = \boldsymbol{J}(m) + \tfrac{\eta}{N}\ \boldsymbol{\xi}(m)\ \mathcal{G}\left[\boldsymbol{J}(m)\cdot\boldsymbol{\xi}(m), \boldsymbol{B}\cdot\boldsymbol{\xi}(m)\right] \\ &\mathrm{batch}: \quad \boldsymbol{J}(m+1) = \boldsymbol{J}(m) + \tfrac{\eta}{N}\ \langle \boldsymbol{\xi}\ \mathcal{G}\left[\boldsymbol{J}(m)\cdot\boldsymbol{\xi}, \boldsymbol{B}\cdot\boldsymbol{\xi}\right]\rangle_{\tilde{D}} \end{aligned} \qquad (2.1)$$

In on-line learning one draws at each iteration step m a question $\boldsymbol{\xi}(m) \in \tilde{D}$ at random, the dynamics is thus a stochastic process; in batch learning one iterates a deterministic map. The function $\mathcal{G}[x,y]$ is assumed to be bounded and not to depend on N, other than via its two arguments.

Our most important observables during learning are the training error $E_{\rm t}(\boldsymbol{J})$ and the generalization error $E_{\rm g}(\boldsymbol{J})$, defined as follows:

$$E_{\rm t}(\boldsymbol{J}) = \langle \theta[-(\boldsymbol{J}\cdot\boldsymbol{\xi})(\boldsymbol{B}\cdot\boldsymbol{\xi})]\rangle_{\tilde{D}} \ , \tag{2.2}$$

$$E_{\rm g}(\boldsymbol{J}) = \langle \theta[-(\boldsymbol{J}\cdot\boldsymbol{\xi})(\boldsymbol{B}\cdot\boldsymbol{\xi})]\rangle_{D} \ . \tag{2.3}$$

Only if the training set $\tilde{D}$ is sufficiently large, and if there are no correlations between $\boldsymbol{J}$ and the questions $\boldsymbol{\xi} \in \tilde{D}$, will these two errors will be identical.

2.2 From Discrete to Continuous Time

We next convert the dynamical laws (2.1) into the language of stochastic processes. We introduce the probability $\hat{p}_m(\boldsymbol{J})$ to find weight vector $\boldsymbol{J}$ at discrete iteration step m. In terms of this microscopic probability distribution the processes (2.1) can be written in the general Markovian form

$$\hat{p}_{m+1}(\boldsymbol{J}) = \int d\boldsymbol{J}' \ W[\boldsymbol{J};\boldsymbol{J}'] \ \hat{p}_m(\boldsymbol{J}') \ ,$$

with the transition probabilities

$$\begin{array}{ll} {\rm on-line:} & W[\boldsymbol{J};\boldsymbol{J}'] = \langle \delta\left[\boldsymbol{J}-\boldsymbol{J}'-\frac{\eta}{N}\,\boldsymbol{\xi}\,\mathcal{G}\left[\boldsymbol{J}'\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}\right]\right]\rangle_{\tilde{D}} \\ {\rm batch:} & W[\boldsymbol{J};\boldsymbol{J}'] = \delta\left[\boldsymbol{J}-\boldsymbol{J}'-\frac{\eta}{N}\langle\boldsymbol{\xi}\,\mathcal{G}\left[\boldsymbol{J}'\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}\right]\rangle_{\tilde{D}}\right] \end{array} \tag{2.4}$$

We now make the transition to a description involving real-valued time labels by choosing the duration of each iteration step to be a real-valued random number, such that the probability that at time t precisely m steps have been made is given by the Poisson expression

$$\pi_m(t) = \frac{1}{m!}(Nt)^m e^{-Nt} \ . \tag{2.5}$$

For times $t \gg N^{-1}$ we find $t = m/N + \mathcal{O}(N^{-\frac{1}{2}})$, the usual time unit. Due to the random durations of the iteration steps we have to switch to the following microscopic probability distribution:

$$p_t(\boldsymbol{J}) = \sum_{m\geq 0} \pi_m(t) \ \hat{p}_m(\boldsymbol{J}) \ .$$

This distribution obeys a simple differential equation, which immediately follows from the pleasant properties of (2.5) under temporal differentiation:

$$\frac{d}{dt}\, p_t(\boldsymbol{J}) = N \int d\boldsymbol{J}' \ \{W[\boldsymbol{J};\boldsymbol{J}'] - \delta[\boldsymbol{J}-\boldsymbol{J}']\} \ p_t(\boldsymbol{J}') \ . \tag{2.6}$$

So far no approximations have been made, equation (2.6) is exact for any N. It is the equivalent of the master equation often introduced to define the dynamics of spin systems.

2.3 Derivation of Macroscopic Fokker-Planck Equation

We now wish to investigate the dynamics of a number of as yet arbitrary *macroscopic* observables $\mathbf{\Omega}[\boldsymbol{J}] = (\Omega_1[\boldsymbol{J}], \ldots, \Omega_k[\boldsymbol{J}])$. To do so we introduce a macroscopic probability distribution

$$P_t(\mathbf{\Omega}) = \int d\boldsymbol{J}\ p_t(\boldsymbol{J})\ \delta\left[\mathbf{\Omega} - \mathbf{\Omega}[\boldsymbol{J}]\right]\ .$$

Its time derivative immediately follows from that in (2.6):

$$\frac{d}{dt}P_t(\mathbf{\Omega}) = \int d\mathbf{\Omega}'\ \mathcal{W}_t[\mathbf{\Omega};\mathbf{\Omega}']\ P_t(\mathbf{\Omega}')\ , \tag{2.7}$$

where

$$\mathcal{W}_t[\mathbf{\Omega};\mathbf{\Omega}'] = \frac{\int d\boldsymbol{J}'\ p_t(\boldsymbol{J}')\ \delta\left[\mathbf{\Omega}'-\mathbf{\Omega}[\boldsymbol{J}']\right]\ \int d\boldsymbol{J}\ \delta\left[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}]\right] N\{W[\boldsymbol{J};\boldsymbol{J}']-\delta[\boldsymbol{J}-\boldsymbol{J}']\}}{\int d\boldsymbol{J}'\ p_t(\boldsymbol{J}')\ \delta\left[\mathbf{\Omega}'-\mathbf{\Omega}[\boldsymbol{J}']\right]}$$

If we insert the relevant expressions (2.4) for $W[\boldsymbol{J};\boldsymbol{J}']$ we can perform the $\boldsymbol{J}$-integrations, and obtain expressions in terms of so-called sub-shell averages, defined as

$$\langle f(\boldsymbol{J})\rangle_{\mathbf{\Omega};t} = \frac{\int d\boldsymbol{J}\ p_t(\boldsymbol{J})\ \delta\left[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}]\right]\ f(\boldsymbol{J})}{\int d\boldsymbol{J}\ p_t(\boldsymbol{J})\ \delta\left[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}]\right]}\ .$$

For the two types of learning rules at hand we obtain:

$$\mathcal{W}_t^{\rm on}[\mathbf{\Omega};\mathbf{\Omega}'] = N\left\langle\!\!\left\langle \delta\left[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}+\frac{\eta}{N}\ \boldsymbol{\xi}\ \mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]]\right]\right\rangle_{\tilde{D}} - \delta\left[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}]\right]\right\rangle_{\mathbf{\Omega}';t}$$

$$\mathcal{W}_t^{\rm ba}[\mathbf{\Omega};\mathbf{\Omega}'] = N\left\langle \delta\left[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}+\frac{\eta}{N}\langle\boldsymbol{\xi}\ \mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\Omega}]\right] - \delta\left[\mathbf{\Omega}-\mathbf{\Omega}[\boldsymbol{J}]\right]\right\rangle_{\mathbf{\Omega}';t}$$

We now insert integral representations for the δ-distributions. This gives for our two learning scenario's:

$$\mathcal{W}_t^{\rm on}[\mathbf{\Omega};\mathbf{\Omega}'] = \int\frac{d\hat{\mathbf{\Omega}}}{(2\pi)^k}e^{i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}}\ N\left\langle\!\!\left\langle e^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}+\frac{\eta}{N}\boldsymbol{\xi}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]]}\right\rangle_{\tilde{D}} - e^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}]}\right\rangle_{\mathbf{\Omega}';t} \tag{2.8}$$

$$\mathcal{W}_t^{\rm ba}[\mathbf{\Omega};\mathbf{\Omega}'] = \int\frac{d\hat{\mathbf{\Omega}}}{(2\pi)^k}e^{i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}}\ N\left\langle e^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}+\frac{\eta}{N}\langle\boldsymbol{\xi}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}}]} - e^{-i\hat{\mathbf{\Omega}}\cdot\mathbf{\Omega}[\boldsymbol{J}]}\right\rangle_{\mathbf{\Omega}';t} \tag{2.9}$$

Still no approximations have been made. The above two expressions differ only in the stage where the averaging over the training set is carried out.

In expanding equations (2.8,2.9) for large N and finite t we have to be careful, since the system size N enters both as a small parameter to control the magnitude of the modification of individual components of the weight vector, but also determines the dimensions and lengths of various vectors that occur. If we assess how derivatives with respect to individual components J_i

scale for observables such as $Q[\boldsymbol{J}]=\boldsymbol{J}^2$ and $R[\boldsymbol{J}]=\boldsymbol{B}\cdot\boldsymbol{J}$, we find the following scaling property which we will choose as our definition of *simple* mean-field observables:

$$F[\boldsymbol{J}]=\mathcal{O}(N^0) \qquad \frac{\partial^\ell F[\boldsymbol{J}]}{\partial J_{i_1}\cdots\partial J_{i_\ell}}=\mathcal{O}(|\boldsymbol{J}|^{-\ell}N^{\frac{1}{2}\ell-d}) \qquad (N\to\infty)\;, \tag{2.10}$$

in which d is the number of *different* elements in the set $\{i_1,\ldots,i_\ell\}$. However, we will find that for restricted training sets not all relevant observables will have the properties (2.10). In particular, the joint distribution of student and teacher fields will, at least for on-line learning, have a contribution for which higher order derivatives do not decrease in importance[1]. The latter type of more *general* mean-field observables will have to be defined via the identities

$$F[\boldsymbol{J}+\boldsymbol{k}]-F[\boldsymbol{J}]=\Delta[\boldsymbol{J};\boldsymbol{k}]+\sum_i k_i\frac{\partial F[\boldsymbol{J}]}{\partial J_i}+\frac{1}{2}\sum_{ij}k_ik_j\frac{\partial^2F[\boldsymbol{J}]}{\partial J_i\partial J_j}+\sum_{\ell\geq 3}\mathcal{O}\left(\frac{|\boldsymbol{k}|^\ell}{|\boldsymbol{J}|^\ell}\right) \tag{2.11}$$

$$F[\boldsymbol{J}]=\mathcal{O}(N^0), \qquad \Delta[\boldsymbol{J};\boldsymbol{k}]=\mathcal{O}(\boldsymbol{k}^2/\boldsymbol{J}^2) \tag{2.12}$$

(in the assessment of the order of the remainder terms of (2.11) we have used $\sum_i k_i=\mathcal{O}(\sqrt{N}|\boldsymbol{k}|)$). Simple mean-field observables correspond to $\Delta[\boldsymbol{J};\boldsymbol{k}]=0$.

We apply (2.11) to our macroscopic equations (2.8,2.9), restricting ourselves from now on to mean-field observables in the sense of (2.11,2.12). One of our observables we choose to be $\boldsymbol{J}^2$. In the present problem the shifts $\boldsymbol{k}$, being either $\frac{\eta}{N}\boldsymbol{\xi}\;\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi};\boldsymbol{B}\cdot\boldsymbol{\xi}]$ or $\frac{\eta}{N}\langle\boldsymbol{\xi}\;\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi};\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}}$, scale as $|\boldsymbol{k}|=\mathcal{O}(N^{-\frac{1}{2}})$. Consequently:

$$e^{-i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}+\boldsymbol{k}]}=e^{-i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}]}\left\{1-\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Delta}[\boldsymbol{J};\boldsymbol{k}]-i\sum_i k_i\frac{\partial}{\partial J_i}(\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}])\right.$$

$$\left.-\frac{i}{2}\sum_{ij}k_ik_j\frac{\partial^2}{\partial J_i\partial J_j}(\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}])-\frac{1}{2}\left[\sum_i k_i\frac{\partial}{\partial J_i}(\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}])\right]^2\right\}+\mathcal{O}(N^{-\frac{3}{2}})\;.$$

This, in turn, gives

$$\int\frac{d\hat{\boldsymbol{\Omega}}}{(2\pi)^k}e^{i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}}\;N\left[e^{-i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}+\boldsymbol{k}]}-e^{-i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}]}\right]$$

$$=-N\left\{\sum_\mu\frac{\partial}{\partial\Omega_\mu}\left[\Delta_\mu[\boldsymbol{J};\boldsymbol{k}]+\sum_i k_i\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}+\frac{1}{2}\sum_{ij}k_ik_j\frac{\partial^2\Omega_\mu[\boldsymbol{J}]}{\partial J_i\partial J_j}\right]\right.$$

$$\left.-\frac{1}{2}\sum_{\mu\nu}\frac{\partial^2}{\partial\Omega_\mu\partial\Omega_\nu}\sum_{ij}k_ik_j\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\frac{\partial\Omega_\nu[\boldsymbol{J}]}{\partial J_j}\right\}\delta\left[\boldsymbol{\Omega}-\boldsymbol{\Omega}[\boldsymbol{J}]\right]+\mathcal{O}(N^{-\frac{1}{2}})\;.$$

[1]We are grateful to Dr. Yuan-sheng Xiong for alerting us to this important point.

It is now evident, in view of (2.8,2.9), that both types of dynamics are described by macroscopic laws with transition probability densities of the general form

$$\mathcal{W}_t^{\star\star}[\mathbf{\Omega};\mathbf{\Omega}'] = \left\{ -\sum_\mu F_\mu[\mathbf{\Omega}';t]\frac{\partial}{\partial\Omega_\mu} + \frac{1}{2}\sum_{\mu\nu} G_{\mu\nu}[\mathbf{\Omega}';t]\frac{\partial^2}{\partial\Omega_\mu\partial\Omega_\nu}\right\}\delta\left[\mathbf{\Omega}-\mathbf{\Omega}'\right] + \mathcal{O}(N^{-\frac{1}{2}})$$

which, due to (2.7) and for $N\to\infty$ and finite times, leads to a Fokker-Planck equation:

$$\frac{d}{dt}P_t(\mathbf{\Omega}) = -\sum_{\mu=1}^{k}\frac{\partial}{\partial\Omega_\mu}\left\{F_\mu[\mathbf{\Omega};t]P_t(\mathbf{\Omega})\right\} + \frac{1}{2}\sum_{\mu\nu=1}^{k}\frac{\partial^2}{\partial\Omega_\mu\partial\Omega_\nu}\left\{G_{\mu\nu}[\mathbf{\Omega};t]P_t(\mathbf{\Omega})\right\}\ . \tag{2.13}$$

The differences between the two types of dynamics are in the explicit expressions for the flow- and diffusion terms:

$$F_\mu^{\rm on}[\mathbf{\Omega};t] = \lim_{N\to\infty}\left\langle N\langle\Delta_\mu[\boldsymbol{J};\frac{\eta}{N}\boldsymbol{\xi}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]]\rangle_{\check{D}} + \eta\sum_i\langle\xi_i\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\check{D}}\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i} + \frac{\eta^2}{2N}\sum_{ij}\langle\xi_i\xi_j\ \mathcal{G}^2[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\check{D}}\frac{\partial^2\Omega_\mu[\boldsymbol{J}]}{\partial J_i\partial J_j}\right\rangle_{\mathbf{\Omega};t}$$

$$G_{\mu\nu}^{\rm on}[\mathbf{\Omega};t] = \lim_{N\to\infty}\frac{\eta^2}{N}\left\langle\sum_{ij}\langle\xi_i\xi_j\ \mathcal{G}^2[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\check{D}}\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\frac{\partial\Omega_\nu[\boldsymbol{J}]}{\partial J_j}\right\rangle_{\mathbf{\Omega};t}$$

$$F_\mu^{\rm ba}[\mathbf{\Omega};t] = \lim_{N\to\infty}\left\langle N\Delta_\mu[\boldsymbol{J};\frac{\eta}{N}\langle\boldsymbol{\xi}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi};\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\check{D}}] + \eta\sum_i\langle\xi_i\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\check{D}}\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i} + \frac{\eta^2}{2N}\sum_{ij}\langle\xi_i\ \mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\check{D}}\langle\xi_j\ \mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\check{D}}\frac{\partial^2\Omega_\mu[\boldsymbol{J}]}{\partial J_i\partial J_j}\right\rangle_{\mathbf{\Omega};t}$$

$$G_{\mu\nu}^{\rm ba}[\mathbf{\Omega};t] = \lim_{N\to\infty}\frac{\eta^2}{N}\left\langle\sum_{ij}\langle\xi_i\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\check{D}}\langle\xi_j\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\check{D}}\frac{\partial\Omega_\mu[\boldsymbol{J}]}{\partial J_i}\frac{\partial\Omega_\nu[\boldsymbol{J}]}{\partial J_j}\right\rangle_{\mathbf{\Omega};t}$$

Equation (2.13) allows us to define the goal of our exercise in more explicit form. If we wish to arrive at closed deterministic macroscopic equations, we have to choose our observables such that

$\lim_{N\to\infty} G_{\mu\nu}[\mathbf{\Omega};t] = 0$ (this ensures determinism)

$\lim_{N\to\infty}\frac{\partial}{\partial t}F_\mu[\mathbf{\Omega};t] = 0$ (this ensures closure)

In the case of time-dependent global parameters, such as learning rates or decay rates, the latter condition relaxes to the requirement that any explicit time-dependence of $F_\mu[\mathbf{\Omega};t]$ is restricted to these global parameters.

3 Application to Canonical Observables

3.1 Choice of Canonical Observables

We now apply the general results obtained so far to a specific set of observables, $\mathbf{\Omega} \to \{Q, R, P\}$, which are taylored to the problem at hand:

$$Q[\boldsymbol{J}] = \boldsymbol{J}^2, \qquad R[\boldsymbol{J}] = \boldsymbol{J}\cdot\boldsymbol{B}, \qquad P[x,y;\boldsymbol{J}] = \langle \delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\ \delta[x-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}} \quad (3.1)$$

with $x, y \in \Re$. This choice is motivated by the following considerations: (i) in order to incorporate the standard theory in the limit $\alpha \to \infty$ we need at least $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$, (ii) we need to be able to calculate the training error, which involves field statistics calculated over the training set $\tilde{D}$, as described by $P[x,y;\boldsymbol{J}]$, and (iii) for finite α one cannot expect closed macroscopic equations for just a finite number of order parameters, the present choice (involving the order parameter *function* $P[x,y;\boldsymbol{J}]$) represents effectively an infinite number[2]. In subsequent calculations we will, however, assume the number of arguments (x,y) for which $P[x,y;\boldsymbol{J}]$ is to be evaluated (and thus our number of order parameters) to go to infinity only after the limit $N \to \infty$ has been taken. This will eliminate many technical subtleties and will allow us to use the Fokker-Planck equation (2.13).

The observables (3.1) are indeed of the general mean-field type in the sense of (2.11,2.12). Insertion into the stronger condition (2.10) immediately shows this to be true for the scalar observables $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$. Verification of (2.11,2.12) for the function $P[x,y;\boldsymbol{J}]$ is less trivial. We denote with $\mathcal{I}$ the set of all *different* indices in the list $(i_1,\ldots,i_\ell)$, with n_k giving the number of times a number k occurs, and with $\mathcal{I}^{\pm} \subseteq \mathcal{I}$ defined as the set of all indices $k \in \mathcal{I}$ for which n_k is even $(+)$, or odd $(-)$. Note that with these definitions $\ell = \sum_{k\in\mathcal{I}^+} n_k + \sum_{k\in\mathcal{I}^-} n_k \geq 2|\mathcal{I}^+| + |\mathcal{I}^-|$. We then have:

$$\frac{\partial^\ell P[x,y;\boldsymbol{J}]}{\partial J_{i_1}\ldots\partial J_{i_\ell}} =$$

$$(-1)^\ell \frac{\partial^\ell}{\partial x^\ell} \int \frac{d\hat{x}d\hat{y}}{(2\pi)^2}\ e^{i[x\hat{x}+y\hat{y}]} \left\langle \left[\prod_{k\in\mathcal{I}} \xi_k^{n_k} e^{-i\xi_k[\hat{x}J_k+\hat{y}B_k]}\right] \left[\prod_{k\notin\mathcal{I}} e^{-i\xi_k[\hat{x}J_k+\hat{y}B_k]}\right] \right\rangle_{\tilde{D}}$$

Upon writing averaging over *all* training sets of size $p = \alpha N$ as $\langle\ldots\rangle_{\mathbf{\Xi}}$, this allows us to conclude

$$\left\langle \frac{\partial^\ell P[x,y;\boldsymbol{J}]}{\partial J_{i_1}\ldots\partial J_{i_\ell}} \right\rangle_{\mathbf{\Xi}} = \mathcal{O}(N^{-\frac{1}{2}|\mathcal{I}^-|})$$

[2] A simple rule of thumb is the following: if a process requires replica theory for its stationary state analysis, as does learning with restricted training sets, its dynamics is of a spin-glass type and cannot be described by a finite set of closed dynamic equations.

Since $\frac{1}{2}\ell-|\mathcal{I}|+\frac{1}{2}|\mathcal{I}^-| = \frac{1}{2}[\ell-|\mathcal{I}^-|-2|\mathcal{I}^+|] \geq 0$, the *average over all training sets* of the function $P[x,y;\boldsymbol{J}]$ is thus found to be a simple mean-field observable.

The scaling properties of expansions or derivatives of $P[x,y;\boldsymbol{J}]$ for a given training set $\tilde{D}$, however, cannot be assumed identical to those of its average over all training sets, due to the statistical dependence of the shifts $\boldsymbol{k}$ in $P[x,y;\boldsymbol{J}+\boldsymbol{k}]$ on the composition of $\tilde{D}$ (such subtleties are absent in the case $\alpha=\infty$ of complete training sets). This dependence turns out to be harmless in the case of batch learning, but will have a considerable impact in the case of on-line learning, where $\boldsymbol{k}^{\rm on}=\eta N^{-1}\boldsymbol{\xi}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]$ is proportional to an individual member of the set $\tilde{D}$. The field distribution $P[x,y;\boldsymbol{J}]$ turns out to obey (2.11,2.12) for both on-line and batch learning (full details can be found in Coolen and Saad, 1998), with

$$\Delta_{xy}[\boldsymbol{J};\boldsymbol{k}^{\rm on}] = \frac{1}{p}\left\{\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}-\eta\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]-\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\right.$$
$$\left.+\eta\frac{\partial}{\partial x}\left[\mathcal{G}[x,y]\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\right]-\frac{1}{2}\eta^2\frac{\partial^2}{\partial x^2}\left[\mathcal{G}^2[x,y]\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\right]\right\} \tag{3.2}$$

For on-line learning the field distribution is apparently not a simple mean-field observable. In contrast $\Delta_{xy}[\boldsymbol{J};\boldsymbol{k}^{\rm ba}]=0$, thus for batch learning the distribution $P[x,y;\boldsymbol{J}]$ is a simple mean field observable in the sense of (2.10). Note that $\Delta_{xy}[\boldsymbol{J};\boldsymbol{k}^{\rm on}]=\mathcal{O}(\eta^3)$ as $\eta\to 0$, so that for small learning rates these differences between batch and on-line learning disappear, as they should. Having chosen our order parameters to be Q, R and $\{P[x,y]\}$, we will from this stage onwards use the notation $\langle\ldots\rangle_{{\rm QRP};t}$ for sub-shell averages defined with respect to this choice.

3.2 Deterministic Dynamical Laws

Here we will first show that for the observables (3.1) the diffusion matrix elements $G^{\star\star}_{\mu\nu}$ in the Fokker-Planck equation (2.13) vanish for $N\to\infty$. Our observables will consequently obey deterministic dynamical laws, which we will calculate in explicit form. We can save ink and trees by introducing the complementary Kronecker delta $\overline{\delta}_{ab}=1-\delta_{ab}$ and the following key functions which we will repeatedly encounter:

$$\mathcal{A}[x,y;x',y'] = \lim_{N\to\infty}$$
$$\left\langle\langle\!\langle\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}(\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\rangle\!\rangle_{\tilde{D}}\right\rangle_{{\rm QRP};t} \tag{3.3}$$
$$\mathcal{B}[x,y;x',y'] = \lim_{N\to\infty}$$
$$\left\langle\frac{1}{N}\sum_{i\neq j}\langle\!\langle\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}(\xi_i\xi_j\xi_i'\xi_j')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\rangle\!\rangle_{\tilde{D}}\right\rangle_{{\rm QRP};t} \tag{3.4}$$

$$\mathcal{C}[x,y;x',y';x'',y''] = \lim_{N\to\infty}\frac{1}{N}\left\langle \langle\!\langle\!\langle\, \overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}''}\overline{\delta}_{\boldsymbol{\xi}'\boldsymbol{\xi}''}(\boldsymbol{\xi}\cdot\boldsymbol{\xi}'')(\boldsymbol{\xi}'\cdot\boldsymbol{\xi}'')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\right.$$
$$\left.\times\,\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\delta[x''-\boldsymbol{J}\cdot\boldsymbol{\xi}'']\delta[y''-\boldsymbol{B}\cdot\boldsymbol{\xi}'']\,\rangle\!\rangle\!\rangle_{\tilde{D}}\right\rangle_{\mathrm{QRP};t} \tag{3.5}$$

$$\mathcal{D}[x,y;x',y';u,v;u',v'] = \lim_{N\to\infty}\frac{1}{N}\left\langle \langle\!\langle\!\langle\!\langle\, \overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}''}\overline{\delta}_{\boldsymbol{\xi}'\boldsymbol{\xi}'''}(\boldsymbol{\xi}\cdot\boldsymbol{\xi}'')(\boldsymbol{\xi}'\cdot\boldsymbol{\xi}''')\right.$$
$$\times\,\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']$$
$$\left.\times\,\delta[u-\boldsymbol{J}\cdot\boldsymbol{\xi}'']\delta[v-\boldsymbol{B}\cdot\boldsymbol{\xi}'']\delta[u'-\boldsymbol{J}\cdot\boldsymbol{\xi}''']\delta[v'-\boldsymbol{B}\cdot\boldsymbol{\xi}''']\rangle\!\rangle\!\rangle\!\rangle_{\tilde{D}}\right\rangle_{\mathrm{QRP};t} \tag{3.6}$$

We show in a subsequent section that, within the present formalism, all three functions (3.4,3.5,3.6) are zero. The function (3.3), on the other hand, will be found not to vanish and to contain all the interesting and non-trivial physics of the process. It plays the role of a Green's function, and its calculation will turn out to be our central problem.

First we turn to the diffusion matrix elements of the macroscopic Fokker-Planck equation (2.13). Calculating the diffusion terms associated with $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$ only is trivial. We write $\langle f[x,y]\rangle = \int dxdy\; P[x,y]f[x,y]$ and find

$$\begin{bmatrix} G^{\mathrm{on}}_{QQ}[\ldots] \\ G^{\mathrm{on}}_{QR}[\ldots] \\ G^{\mathrm{on}}_{RR}[\ldots] \end{bmatrix} = \lim_{N\to\infty}\frac{\eta^2}{N}\begin{bmatrix} 4\langle x^2\mathcal{G}^2[x,y]\rangle \\ 2\langle xy\mathcal{G}^2[x,y]\rangle \\ \langle y^2\mathcal{G}^2[x,y]\rangle \end{bmatrix} = 0$$

$$\begin{bmatrix} G^{\mathrm{ba}}_{QQ}[\ldots] \\ G^{\mathrm{ba}}_{QR}[\ldots] \\ G^{\mathrm{ba}}_{RR}[\ldots] \end{bmatrix} = \lim_{N\to\infty}\frac{\eta^2}{N}\begin{bmatrix} 4\langle x\mathcal{G}[x,y]\rangle^2 \\ 2\langle x\mathcal{G}[x,y]\rangle\langle y\mathcal{G}[x,y]\rangle \\ \langle y\mathcal{G}[x,y]\rangle^2 \end{bmatrix} = 0$$

In calculating diffusion terms which involve the function $P[x;y;\boldsymbol{J}]$ we will need two simple scaling consequences of the random composition of $\tilde{D}$:

$$\boldsymbol{\xi}\in\tilde{D}:\ \sum_{\boldsymbol{\xi}'\in\tilde{D}}\delta_{\boldsymbol{\xi}\boldsymbol{\xi}'} = 1+\mathcal{O}(N^{-1}) \qquad \text{and} \qquad \frac{1}{p^2}\sum_{\boldsymbol{\xi},\boldsymbol{\xi}'\in\tilde{D}}\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}|\boldsymbol{\xi}\cdot\boldsymbol{\xi}'| = \mathcal{O}(N^{\frac{1}{2}})$$

For the diffusion terms with just one occurrence of $P[x,y]$ we now find:

$$\begin{bmatrix} G^{\mathrm{on}}_{Q,P[x,y]}[\ldots] \\ G^{\mathrm{on}}_{R,P[x,y]}[\ldots] \end{bmatrix} = -\eta^2\int dx'dy'\;\mathcal{G}^2[x',y']\begin{bmatrix}2x'\\ y'\end{bmatrix}\frac{\partial}{\partial x}\Big\{$$
$$\lim_{N\to\infty}\frac{1}{N}\left\langle \langle\!\langle(\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\rangle\!\rangle_{\tilde{D}}\right\rangle_{\mathrm{QRP};t}\Big\}$$
$$= -\eta^2\int dx'dy'\;\mathcal{G}^2[x',y']\begin{bmatrix}2x'\\ y'\end{bmatrix}\frac{\partial}{\partial x}\left\{\mathcal{O}(N^{-\frac{1}{2}})\right\} = 0$$

$$\left[\begin{array}{c} G^{\rm ba}_{Q,P[x,y]}[\ldots] \\ G^{\rm ba}_{R,P[x,y]}[\ldots] \end{array}\right] = -\eta^2 \left[\begin{array}{c} \langle 2x\mathcal{G}[x,y]\rangle \\ \langle y\mathcal{G}[x,y]\rangle \end{array}\right] \int dx'dy'\; \mathcal{G}[x',y'] \frac{\partial}{\partial x} \Big\{$$

$$\lim_{N\to\infty} \frac{1}{N} \Big\langle \langle\!\langle (\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\rangle\!\rangle_{\tilde{D}} \Big\rangle_{\rm QRP;t} \Big\}$$

$$= -\eta^2 \left[\begin{array}{c} \langle 2x\mathcal{G}[x,y]\rangle \\ \langle y\mathcal{G}[x,y]\rangle \end{array}\right] \int dx'dy'\; \mathcal{G}[x',y'] \frac{\partial}{\partial x} \Big\{ \mathcal{O}(N^{-\frac{1}{2}}) \Big\} = 0$$

The non-trivial terms are those where two derivatives of the function $P[x,y;\boldsymbol{J}]$ come into play. Here we must separate four distinct contributions, defined according to which of the vectors from the trio $\{\boldsymbol{\xi},\boldsymbol{\xi}',\boldsymbol{\xi}''\}$ are identical:

$$G^{\rm on}_{P[x,y],P[x',y']}[\ldots] = \eta^2 \int dx''dy'' \mathcal{G}^2[x'',y''] \frac{\partial^2}{\partial x\partial x'} \lim_{N\to\infty} \frac{1}{N} \Big\langle \langle\!\langle\!\langle (\boldsymbol{\xi}\cdot\boldsymbol{\xi}'')(\boldsymbol{\xi}'\cdot\boldsymbol{\xi}'')\times$$

$$\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\delta[x''-\boldsymbol{J}\cdot\boldsymbol{\xi}'']\delta[y''-\boldsymbol{B}\cdot\boldsymbol{\xi}'']\rangle\!\rangle\!\rangle_{\tilde{D}} \Big\rangle_{\rm QRP;t}$$

$$= \eta^2 \int dx''dy''\; \mathcal{G}^2[x'',y''] \frac{\partial^2}{\partial x\partial x'} \Big\{ \mathcal{C}[x,y;x',y';x'',y'']$$

$$+ \Big[\delta[x''-x]\delta[y''-y] + \delta[x''-x']\delta[y''-y']\Big] \lim_{N\to\infty} \mathcal{O}(N^{-\frac{1}{2}})$$

$$+ \delta[x''-x]\delta[y''-y]\delta[x'-x]\delta[y'-y] \lim_{N\to\infty} \mathcal{O}(N^{-1}) \Big\}$$

$$= \eta^2 \int dx''dy''\; \mathcal{G}^2[x'',y''] \frac{\partial^2}{\partial x\partial x'} \mathcal{C}[x,y;x',y';x'',y'']$$

Similarly:

$$G^{\rm ba}_{P[x,y],P[x',y']}[\ldots] = \eta^2 \int dudvdu'dv'\; \mathcal{G}[u,v]\mathcal{G}[u',v'] \frac{\partial^2}{\partial x\partial x'}$$

$$\lim_{N\to\infty} \frac{1}{N} \Big\langle \langle\!\langle (\boldsymbol{\xi}\cdot\boldsymbol{\xi}'')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[u-\boldsymbol{J}\cdot\boldsymbol{\xi}'']\delta[v-\boldsymbol{B}\cdot\boldsymbol{\xi}'']\rangle\!\rangle_{\tilde{D}}$$

$$\times \langle\!\langle (\boldsymbol{\xi}'\cdot\boldsymbol{\xi}'')\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\delta[u'-\boldsymbol{J}\cdot\boldsymbol{\xi}'']\delta[v'-\boldsymbol{B}\cdot\boldsymbol{\xi}'']\rangle\!\rangle_{\tilde{D}} \Big\rangle_{\rm QRP;t}$$

$$= \eta^2 \int dudvdu'dv'\; \mathcal{G}[u,v]\mathcal{G}[u',v'] \frac{\partial^2}{\partial x\partial x'} \mathcal{D}[x,y;x',y';u,v;u',v']$$

Anticipating the two functions $\mathcal{C}[\ldots]$ and $\mathcal{D}[\ldots]$ of (3.5,3.6) to be zero (to be demonstrated in a subsequent section) we conclude that all diffusion terms vanish. The macroscopic Fokker-Planck equation (2.13) thereby reduces to a Liouville equation, describing deterministic evolution for our macroscopic observables: $\frac{d}{dt}\boldsymbol{\Omega} = \boldsymbol{F}[\boldsymbol{\Omega};t]$. These deterministic equations we will now work out explicitly.

On-Line Learning

First we deal with the scalar observables Q and R, whose equations are worked out easily to give

$$\frac{d}{dt}Q = 2\eta \int dxdy\ P[x,y]\ x\mathcal{G}[x,y] + \eta^2 \int dxdy\ P[x,y]\ \mathcal{G}^2[x,y] \tag{3.7}$$

$$\frac{d}{dt}R = \eta \int dxdy\ P[x,y]\ y\mathcal{G}[x,y] \tag{3.8}$$

These equations are identical to those of the familiar $\alpha \to \infty$ formalism. The difference is in the function to be substituted for $P[x,y]$, which would have been a simple Gaussian one for $\alpha \to \infty$, but which here is the solution of

$$\frac{\partial}{\partial t}P[x,y] = \frac{1}{\alpha}\left[\int dx' P[x',y]\delta[x-x'-\eta\mathcal{G}[x',y]] - P[x,y]\right]$$

$$-\eta\frac{\partial}{\partial x}\int dx'dy'\ \mathcal{G}[x',y']\mathcal{A}[x,y;x',y']$$

$$+\frac{1}{2}\eta^2\int dx'dy'\mathcal{G}^2[x',y']P[x',y']\frac{\partial^2}{\partial x^2}P[x,y] + \frac{1}{2}\eta^2\frac{\partial^2}{\partial x^2}\int dx'dy'\mathcal{G}^2[x',y']\mathcal{B}[x,y;x',y']$$

Anticipating the term $\mathcal{B}[\ldots]$ as defined in (3.4) to be zero (to be demonstrated in a subsequent section) we thus arrive at the following compact result:

$$\frac{\partial}{\partial t}P[x,y] = \frac{1}{\alpha}\left[\int dx' P[x',y]\delta[x-x'-\eta\mathcal{G}[x',y]] - P[x,y]\right]$$

$$-\eta\frac{\partial}{\partial x}\int dx'dy'\ \mathcal{G}[x',y']\mathcal{A}[x,y;x',y'] + \frac{1}{2}\eta^2\int dx'dy'\ \mathcal{G}^2[x',y']P[x',y']\frac{\partial^2}{\partial x^2}P[x,y] \tag{3.9}$$

Batch Learning

For Q and R we again find simple and transparent equations:

$$\frac{d}{dt}Q = 2\eta \int dxdy\ P[x,y]\ x\mathcal{G}[x,y] \tag{3.10}$$

$$\frac{d}{dt}R = \eta \int dxdy\ P[x,y]\ y\mathcal{G}[x,y] \tag{3.11}$$

Finally we calculate for batch learning the temporal derivative of the joint field distribution:

$$\frac{\partial}{\partial t}P[x,y] = -\frac{\eta}{\alpha}\frac{\partial}{\partial x}\left[\mathcal{G}[x,y]P[x,y]\right] - \eta\frac{\partial}{\partial x}\int dx'dy'\ \mathcal{A}[x,y;x',y']\mathcal{G}[x',y']$$

$$+\frac{1}{2}\eta^2\frac{\partial^2}{\partial x^2}\int dx'dy'dx''dy''\mathcal{C}[x,y;x',y';x'',y'']\mathcal{G}[x',y']\mathcal{G}[x'',y']$$

Anticipating the term $\mathcal{C}[\ldots]$ as defined in (3.5) to be zero (to be demonstrated in a subsequent section) we thus arrive at the following compact result:

$$\frac{\partial}{\partial t}P[x,y] = -\frac{\eta}{\alpha}\frac{\partial}{\partial x}\Big[\mathcal{G}[x,y]P[x,y]\Big] - \eta\frac{\partial}{\partial x}\int dx'dy'\ \mathcal{A}[x,y;x',y']\mathcal{G}[x',y'] \tag{3.12}$$

Comparing (3.7-3.9) with (3.10-3.12) shows that, as for complete training sets (Mace and Coolen, 1998), the difference between the macroscopic laws for batch and on-line learning is merely the presence (on-line) or absence (batch) of terms which are not linear in the learning rate η (i.e. of order η^2 or higher). This is consistent with the picture that for sufficiently small learning rates the differences between batch and on-line learning must vanish.

3.3 Closure of Macroscopic Dynamical Laws

We close our macroscopic laws (for on-line and batch learning) by making, for $N\to\infty$, the two key assumptions underlying dynamical replica theories:

1. *The observables $\{Q,R,P\}$ obey closed macroscopic dynamic equations.*
2. *These macroscopic dynamic equations are self-averaging with respect to the disorder, i.e. the microscopic realization of the training set $\tilde{D}$.*

Assumption 1 implies that all microscopic probability variations within the $\{Q,R,P\}$ subshells of the $\boldsymbol{J}$-ensemble are either absent or irrelevant to the evolution of $\{Q,R,P\}$. We may consequently make the simplest self-consistent choice for $p_t(\boldsymbol{J})$ in evaluating the macroscopic laws, i.e. in (3.3): microscopic probability equipartitioning in the $\{Q,R,P\}$-subshells of the ensemble, or

$$p_t(\boldsymbol{J})\ \to\ w(\boldsymbol{J}) \sim \delta[Q-Q[\boldsymbol{J}]]\ \delta[R-R[\boldsymbol{J}]]\ \prod_{xy}\delta[P[x,y]-P[x,y;\boldsymbol{J}]] \tag{3.13}$$

The new distribution $w(\boldsymbol{J})$ depends on time only via $\{Q,R,P\}$. Note that (3.13) leads to exact macroscopic laws if for $N\to\infty$ our observables $\{Q,R,P\}$ indeed obey closed equations, and is true in equilibrium for detailed balance models in which the Hamiltonian can be written in terms of $\{Q,R,P\}$. It is an approximation if our observables do not obey closed equations. Assumption 2 allows us to average the macroscopic laws over the disorder; for mean-field models it is usually convincingly supported by numerical simulations, and can be proven within the path integral formalism (see e.g. Horner, 1992). We write averages over all training sets $\tilde{D}\subseteq\{-1,1\}^N$ of size p as $\langle\ldots\rangle_{\Xi}$. Our assumptions result in the closure of both (3.7-3.9) and (3.10-3.12), since now the function $\mathcal{A}[x,y;x',y']$ of (3.3) is expressed fully in terms of $\{Q,R,P\}$:

$$\mathcal{A}[x,y;x',y'] = \lim_{N\to\infty}$$

$$\left\langle \frac{\int d\boldsymbol{J}\ w(\boldsymbol{J})\ \langle\!\langle \delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\rangle\!\rangle_{\tilde{D}}}{\int d\boldsymbol{J}\ w(\boldsymbol{J})} \right\rangle_{\Xi}$$

The final ingredient of dynamical replica theory is the realization that averages of fractions can be calculated with the replica identity

$$\left\langle \frac{\int d\boldsymbol{J}\; W[\boldsymbol{J},\boldsymbol{z}]G[\boldsymbol{J},\boldsymbol{z}]}{\int d\boldsymbol{J}\; W[\boldsymbol{J},\boldsymbol{z}]} \right\rangle_{\boldsymbol{z}} = \lim_{n\to 0} \int d\boldsymbol{J}^1 \cdots d\boldsymbol{J}^n\; \langle G[\boldsymbol{J}^1,\boldsymbol{z}] \prod_{\alpha=1}^{n} W[\boldsymbol{J}^\alpha,\boldsymbol{z}]\rangle_{\boldsymbol{z}}$$

giving

$$\mathcal{A}[x,y;x',y'] = \lim_{N\to\infty}\lim_{n\to 0} \int \prod_{\alpha=1}^{n} w(\boldsymbol{J}^\alpha)\; d\boldsymbol{J}^\alpha$$
$$\left\langle \langle\!\langle \delta[x-\boldsymbol{J}^1\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}\delta[x'-\boldsymbol{J}^1\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\rangle\!\rangle_{\tilde{D}} \right\rangle_{\Xi}$$

Since each weight component scales as $J_i^\alpha = \mathcal{O}(N^{-\frac{1}{2}})$ we transform variables in such a way that our calculations will involve $\mathcal{O}(1)$ objects:

$$(\forall i)(\forall\alpha): \qquad J_i^\alpha = (Q/N)^{\frac{1}{2}}\sigma_i^\alpha, \qquad B_i = N^{-\frac{1}{2}}\tau_i$$

This ensures $\sigma_i^\alpha = \mathcal{O}(1)$, $\tau_i = \mathcal{O}(1)$, and reduces various constraints to ordinary spherical ones: $(\boldsymbol{\sigma}^\alpha)^2 = \boldsymbol{\tau}^2 = N$ for all α. Overall prefactors generated by these transformations always vanish due to $n \to 0$. We find a new effective measure: $\prod_{\alpha=1}^n w(\boldsymbol{J}^\alpha)\; d\boldsymbol{J}^\alpha \to \prod_{\alpha=1}^n \tilde{w}(\boldsymbol{\sigma}^\alpha)\; d\boldsymbol{\sigma}^\alpha$, with

$$\tilde{w}(\boldsymbol{\sigma}) \sim \delta\left[N-\boldsymbol{\sigma}^2\right]\delta\left[NRQ^{-\frac{1}{2}}-\boldsymbol{\tau}\cdot\boldsymbol{\sigma}\right]\prod_{xy}\delta\left[P[x,y]-P[x,y;(Q/N)^{\frac{1}{2}}\boldsymbol{\sigma}]\right] \tag{3.14}$$

In the same fashion one can also express $P[x,y]$ in replica form (which will prove useful for normalization purposes and for self-consistency tests). We thus arrive at

$$\mathcal{A}[x,y;x',y'] = \lim_{n\to 0}\lim_{N\to\infty}\int\prod_{\alpha=1}^{n}\tilde{w}(\boldsymbol{\sigma}^\alpha)\,d\boldsymbol{\sigma}^\alpha \left\langle \left\langle\!\!\left\langle (\boldsymbol{\xi}'\cdot\boldsymbol{\xi})\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}\right.\right.\right.$$
$$\left.\left.\left. \times\,\delta\left[x-\frac{\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]\delta\left[y-\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]\delta\left[x'-\frac{\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right]\delta\left[y-\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right]\right\rangle\!\!\right\rangle_{\tilde{D}}\right\rangle_{\Xi} \tag{3.15}$$

and

$$P_t[x,y] = \lim_{n\to 0}\lim_{N\to\infty}\int\prod_{\alpha=1}^{n}\tilde{w}(\boldsymbol{\sigma}^\alpha)d\boldsymbol{\sigma}^\alpha\left\langle\left\langle\delta\left[x-\frac{\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]\delta\left[y-\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]\right\rangle_{\tilde{D}}\right\rangle_{\Xi} \tag{3.16}$$

Similarly we find replica expressions for the three functions $\mathcal{B}[\ldots]$, $\mathcal{C}[\ldots]$ and $\mathcal{D}[\ldots]$ (3.4-3.6), which will be used subsequently to demonstrate that, within the present formalism, they are self-consistently found to be zero.

We have now converted our problem from a conceptual one into a purely technical one: the evaluation of the integrals and averages in (3.15,3.16), and in similar expressions found for $\mathcal{B}[\ldots]$, $\mathcal{C}[\ldots]$ and $\mathcal{D}[\ldots]$, and the subsequent solution of the resulting closed macroscopic dynamical laws (3.7-3.9) (on-line) and (3.10-3.12) (batch) for the order parameters $\{Q,R,P\}$.

4 Replica Calculation of the Green's Function

4.1 Disorder Averaging

In order to perform the disorder average we insert integral representations for the δ-functions which define the fields (x, y, x', y') and for the δ-functions in the measure (3.14) which involve $P[x, y]$, generating n conjugate order parameter functions $\hat{P}_\alpha(x, y)$. Upon also writing averages over the training set in terms of the p constituent vectors $\{\boldsymbol{\xi}^\mu\}$ we obtain for (3.15) and (3.16):

$$\mathcal{A}[x,y;x',y'] = \int \frac{d\hat{x}\ d\hat{x}' d\hat{y}\ d\hat{y}'}{(2\pi)^4} e^{i[x\hat{x}+x'\hat{x}'+y\hat{y}+y\hat{y}']} \lim_{n\to 0}\lim_{N\to\infty} \int \prod_{\alpha=1}^{n} \prod_{x''y''} d\hat{P}_\alpha(x'',y'')$$

$$\int \prod_{\alpha=1}^{n} \left\{ d\boldsymbol{\sigma}^\alpha\ \delta\left[N-(\boldsymbol{\sigma}^\alpha)^2\right] \delta\left[\frac{NR}{\sqrt{Q}} - \boldsymbol{\tau}\cdot\boldsymbol{\sigma}^\alpha\right] e^{iN\int dx''dy''\ \hat{P}_\alpha(x'',y'')\ P_t(x'',y'')} \right\}$$

$$\left\langle \frac{1}{p^2}\sum_{\mu\neq\nu}(\boldsymbol{\xi}^\mu\boldsymbol{\xi}^\nu) e^{-\frac{i}{\alpha}\sum_{\alpha\lambda}\hat{P}_\alpha\left(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}, \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}\right) - \frac{i}{\sqrt{N}}\boldsymbol{\xi}^\mu\cdot[\hat{x}\sqrt{Q}\boldsymbol{\sigma}^1+\hat{y}\boldsymbol{\tau}] - \frac{i}{\sqrt{N}}\boldsymbol{\xi}^\nu\cdot[\hat{x}'\sqrt{Q}\boldsymbol{\sigma}^1+\hat{y}'\boldsymbol{\tau}]} \right\rangle_{\Xi} \tag{4.1}$$

$$P[x,y] = \int \frac{d\hat{x}\ d\hat{y}}{(2\pi)^2} e^{i[x\hat{x}+y\hat{y}]} \lim_{n\to 0}\lim_{N\to\infty} \int \prod_{\alpha=1}^{n} \prod_{x''y''} d\hat{P}_\alpha(x'',y'')$$

$$\int \prod_{\alpha=1}^{n} \left\{ d\boldsymbol{\sigma}^\alpha \delta\left[N-(\boldsymbol{\sigma}^\alpha)^2\right] \delta\left[\frac{NR}{\sqrt{Q}} - \boldsymbol{\tau}\cdot\boldsymbol{\sigma}^\alpha\right] e^{iN\int dx''dy''\ \hat{P}_\alpha(x'',y'')\ P_t(x'',y'')} \right\}$$

$$\left\langle \frac{1}{p}\sum_{\mu=1}^{p} e^{-\frac{i}{\alpha}\sum_{\alpha\lambda}\hat{P}_\alpha\left(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}, \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}\right) - \frac{i}{\sqrt{N}}\boldsymbol{\xi}^\mu\cdot[\hat{x}\sqrt{Q}\boldsymbol{\sigma}^1+\hat{y}\boldsymbol{\tau}]} \right\rangle_{\Xi} \tag{4.2}$$

In calculating the averages over the training sets $\langle \ldots \rangle_{\Xi}$ that occur in (4.1) and (4.2) one can use permutation symmetries with respect to sites and pattern labels, leading to the following compact results:

$$\left\langle \frac{1}{p^2}\sum_{\mu\neq\nu}(\boldsymbol{\xi}^\mu\boldsymbol{\xi}^\nu) e^{-\frac{i}{\alpha}\sum_{\alpha}\sum_{\lambda}\hat{P}_\alpha\left(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}, \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}\right) - \frac{i}{\sqrt{N}}\boldsymbol{\xi}^\mu\cdot[\hat{x}\sqrt{Q}\boldsymbol{\sigma}^1+\hat{y}\boldsymbol{\tau}] - \frac{i}{\sqrt{N}}\boldsymbol{\xi}^\nu\cdot[\hat{x}'\sqrt{Q}\boldsymbol{\sigma}^1+\hat{y}'\boldsymbol{\tau}]} \right\rangle_{\Xi}$$

$$= e^{p\log\mathcal{D}[0,0]}\ \frac{1}{N}\sum_j \frac{\mathcal{E}_j[\hat{x},\hat{y}]\mathcal{E}_j[\hat{x}',\hat{y}']}{\mathcal{D}^2[0,0]} + \mathcal{O}(N^{-\frac{1}{2}}) \tag{4.3}$$

and

$$\left\langle \frac{1}{p}\sum_{\mu=1}^{p} e^{-\frac{i}{\alpha}\sum_{\alpha}\sum_{\lambda}\hat{P}_\alpha\left(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}, \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}\right) - \frac{i}{\sqrt{N}}\boldsymbol{\xi}^\mu\cdot[\hat{x}\sqrt{Q}\boldsymbol{\sigma}^1+\hat{y}\boldsymbol{\tau}]} \right\rangle_{\Xi}$$

$$= e^{p\log\mathcal{D}[0,0]}\frac{\mathcal{D}[\hat{x},\hat{y}]}{\mathcal{D}[0,0]} + \mathcal{O}\left(N^{-\frac{1}{2}}\right) \tag{4.4}$$

in which

$$\mathcal{D}[u,v] = \left\langle e^{-\frac{i}{\alpha}\sum_\alpha \hat{P}_\alpha\left(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right)-\frac{i}{\sqrt{N}}\boldsymbol{\xi}\cdot[u\sqrt{Q}\boldsymbol{\sigma}^1+v\boldsymbol{\tau}]}\right\rangle_{\boldsymbol{\xi}}$$

$$\mathcal{E}_j[u,v] = \left\langle \sqrt{N}\xi_j\; e^{-\frac{i}{\alpha}\sum_\alpha \hat{P}_\alpha\left(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right)-\frac{i}{\sqrt{N}}\boldsymbol{\xi}\cdot[u\sqrt{Q}\boldsymbol{\sigma}^1+v\boldsymbol{\tau}]}\right\rangle_{\boldsymbol{\xi}}$$

and with the abbreviation $\langle f[\boldsymbol{\xi}]\rangle_{\boldsymbol{\xi}} = 2^{-N}\sum_{\boldsymbol{\xi}\in\{-1,1\}^N} f[\boldsymbol{\xi}]$. These quantities (which are both $\mathcal{O}(1)$ for $N\to\infty$) are, in turn, evaluated by using the central limit theorem, which ensures that for $N\to\infty$ the n rescaled inner products $\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}/\sqrt{N}$ and the rescaled inner product $\boldsymbol{\tau}\cdot\boldsymbol{\xi}/\sqrt{N}$ will become (correlated) zero-average Gaussian variables. After some algebra one finds

$$\begin{aligned}
\mathcal{L}[u,v;u',v'] &= \frac{1}{N}\sum_j \mathcal{E}_j[u,v]\,\mathcal{E}_j[u',v'] = \\
&= -Q\sum_{\alpha\beta} q_{\alpha\beta}(\{\boldsymbol{\sigma}\})\left[\frac{1}{\alpha}\mathcal{F}_1^\alpha[u,v]+u\;\delta_{\alpha 1}\mathcal{D}[u,v]\right]\left[\frac{1}{\alpha}\mathcal{F}_1^\beta[u',v']+u'\;\delta_{\beta 1}\mathcal{D}[u',v']\right] \\
&\quad -R\sum_{\alpha\beta}\left[\frac{1}{\alpha}\mathcal{F}_1^\alpha[u,v]+u\;\delta_{\alpha 1}\mathcal{D}[u,v]\right]\left[\frac{1}{\alpha}\mathcal{F}_2^\beta[u',v']+v'\;\delta_{\beta 1}\mathcal{D}[u',v']\right] \\
&\quad -R\sum_{\alpha\beta}\left[\frac{1}{\alpha}\mathcal{F}_1^\alpha[u',v']+u'\;\delta_{\alpha 1}\mathcal{D}[u',v']\right]\left[\frac{1}{\alpha}\mathcal{F}_2^\beta[u,v]+v\;\delta_{\beta 1}\mathcal{D}[u,v]\right] \\
&\quad -\sum_{\alpha\beta}\left[\frac{1}{\alpha}\mathcal{F}_2^\alpha[u,v]+v\;\delta_{\alpha 1}\mathcal{D}[u,v]\right]\left[\frac{1}{\alpha}\mathcal{F}_2^\beta[u',v']+v'\;\delta_{\beta 1}\mathcal{D}[u',v']\right]+\mathcal{O}(N^{-\frac{1}{2}})
\end{aligned} \tag{4.5}$$

in which $\mathcal{D}[u,v]$ and the $\mathcal{F}_\lambda^\alpha[u,v]$ are given by $n+1$ dimensional integrals:

$$\mathcal{D}[u,v] = \int\frac{d\boldsymbol{x}\;dy\;\det^{\frac{1}{2}}\boldsymbol{A}}{(2\pi)^{(n+1)/2}}\;e^{-\frac{1}{2}\begin{pmatrix}\boldsymbol{x}\\ y\end{pmatrix}\cdot\boldsymbol{A}\begin{pmatrix}\boldsymbol{x}\\ y\end{pmatrix}-\frac{i}{\alpha}\sum_\alpha\hat{P}_\alpha(\sqrt{Q}x_\alpha,y)-i[u\sqrt{Q}x_1+vy]} \tag{4.6}$$

$$\mathcal{F}_\lambda^\alpha[u,v] = \int\frac{d\boldsymbol{x}\;dy\;\det^{\frac{1}{2}}\boldsymbol{A}}{(2\pi)^{(n+1)/2}}\;\partial_\lambda\hat{P}_\alpha(\sqrt{Q}x_\alpha,y)\;e^{-\frac{1}{2}\begin{pmatrix}\boldsymbol{x}\\ y\end{pmatrix}\cdot\boldsymbol{A}\begin{pmatrix}\boldsymbol{x}\\ y\end{pmatrix}-\frac{i}{\alpha}\sum_\alpha\hat{P}_\alpha(\sqrt{Q}x_\alpha,y)-i[u\sqrt{Q}x_1+vy]} \tag{4.7}$$

with $\lambda\in\{1,2\}$. The matrix $\boldsymbol{A}$ in (4.6,4.7) is defined by

$$\boldsymbol{A}^{-1} = \begin{pmatrix} q_{11} & \cdots & q_{1n} & R/\sqrt{Q}\\ \vdots & & \vdots & \vdots\\ q_{n1} & \cdots & q_{nn} & R/\sqrt{Q}\\ R/\sqrt{Q} & \cdots & R/\sqrt{Q} & 1\end{pmatrix}\qquad q_{\alpha\beta}(\{\boldsymbol{\sigma}\}) = \frac{1}{N}\sum_i\sigma_i^\alpha\sigma_i^\beta \tag{4.8}$$

Note that the quantities (4.6,4.7) depend on the microscopic variables $\boldsymbol{\sigma}^\alpha$ only through the spin-glass order parameters $q_{\alpha\beta}(\{\boldsymbol{\sigma}\})$.

It is a straightforward exercise to carry out a similar calculation for the functions $\mathcal{B}[\ldots]$ (3.4), $\mathcal{C}[\ldots]$ (3.4) and $\mathcal{D}[\ldots]$. This gives a zero result in all three cases, basically due to the three functions involving too many unpaired pattern components (each of which will effectively generate a factor $N^{-\frac{1}{2}}$), confirming our previous assertions about the vanishing of all diffusion matrix elements in the macroscopic Fokker-Planck equation (2.13). For details we refer to (Coolen and Saad, 1998).

4.2 Derivation of Saddle-Point Equations

We combine the results (4.3,4.4,4.5) with (4.1,4.2). We use integral representations for the remaining δ-functions, and isolate the $q_{\alpha\beta}$, by inserting

$$1 = \int \frac{d\boldsymbol{q}\; d\hat{\boldsymbol{q}}\; d\hat{\boldsymbol{Q}}\; d\hat{\boldsymbol{R}}}{(2\pi/N)^{n^2+2n}}\; e^{iN[\sum_\alpha(\hat{Q}_\alpha+\hat{R}_\alpha R/\sqrt{Q})+\sum_{\alpha\beta}\hat{q}_{\alpha\beta}q_{\alpha\beta}]}$$
$$\times\; e^{-i\sum_i[\sum_\alpha(\hat{Q}_\alpha(\sigma_i^\alpha)^2+\hat{R}_\alpha\tau_i\sigma_i^\alpha)-i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_i^\alpha\sigma_i^\beta]}$$

We hereby achieve a full factorization over sites, and both (4.1) and (4.2) can be written in the form of an integral dominated by saddle-points:

$$\mathcal{A}[x,y;x',y'] = \int\frac{d\hat{x}\; d\hat{x}'d\hat{y}\; d\hat{y}'}{(2\pi)^4}\; e^{i[x\hat{x}+x'\hat{x}'+y\hat{y}+y\hat{y}']}$$
$$\lim_{n\to0}\lim_{N\to\infty}\int d\boldsymbol{q}\; d\hat{\boldsymbol{q}}\; d\hat{\boldsymbol{Q}}\; d\hat{\boldsymbol{R}}\prod_{\alpha x''y''}d\hat{P}_\alpha(x'',y'')\; e^{N\Psi[\boldsymbol{q},\hat{\boldsymbol{q}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{R}},\{\hat{P}\}]}\frac{\mathcal{L}[\hat{x},\hat{y};\hat{x}',\hat{y}']}{\mathcal{D}^2[0,0]}$$

$$P[x,y] = \int\frac{d\hat{x}\; d\hat{y}}{(2\pi)^2}\; e^{i[x\hat{x}+y\hat{y}]}$$
$$\lim_{n\to0}\lim_{N\to\infty}\int d\boldsymbol{q}\; d\hat{\boldsymbol{q}}\; d\hat{\boldsymbol{Q}}\; d\hat{\boldsymbol{R}}\prod_{\alpha x''y''}d\hat{P}_\alpha(x'',y'')\; e^{N\Psi[\boldsymbol{q},\hat{\boldsymbol{q}},\hat{\boldsymbol{Q}},\hat{\boldsymbol{R}},\{\hat{P}\}]}\frac{\mathcal{D}[\hat{x},\hat{y}]}{\mathcal{D}[0,0]}$$

with

$$\Psi[\ldots] = i\sum_\alpha(\hat{Q}_\alpha+\hat{R}_\alpha R/\sqrt{Q})+i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\;q_{\alpha\beta}+i\sum_\alpha\int dx\;dy\;\hat{P}_\alpha(x,y)P[x,y]$$
$$+\alpha\log\mathcal{D}[0,0]+\lim_{N\to\infty}\frac{1}{N}\sum_i\log\int d\boldsymbol{\sigma}\; e^{-i\sum_\alpha[\hat{Q}_\alpha\sigma_\alpha^2+\hat{R}_\alpha\tau_i\sigma_\alpha]-i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta}$$

The above expressions for $\mathcal{A}[x,y;x',y']$ and $P[x,y]$ will be given by the intensive parts of the integrands, evaluated in the dominating saddle-point of Ψ. We can use the equation for $P[x,y]$ to verify that all expressions are properly normalized. After a simple transformation of some integration variables,

$$\hat{q}_{\alpha\beta}\to\hat{q}_{\alpha\beta}-\hat{Q}_\alpha\delta_{\alpha\beta}\qquad\qquad\hat{R}_\alpha\to\sqrt{Q}\hat{R}_\alpha$$

we arrive at the simple result

$$\mathcal{A}[x,y;x',y'] = \int \frac{d\hat{x}\ d\hat{x}'\ d\hat{y}\ d\hat{y}'}{(2\pi)^4} e^{i[x\hat{x}+x'\hat{x}'+y\hat{y}+y\hat{y}']} \lim_{n\to 0} \frac{\mathcal{L}[\hat{x},\hat{y};\hat{x}',\hat{y}']}{\mathcal{D}^2[0,0]} \quad (4.9)$$

$$P[x,y] = \int \frac{d\hat{x}\ d\hat{y}}{(2\pi)^2} e^{i[x\hat{x}+y\hat{y}]} \lim_{n\to 0} \frac{\mathcal{D}[\hat{x},\hat{y}]}{\mathcal{D}[0,0]} \quad (4.10)$$

in which all functions are to be evaluated upon choosing for the order parameters the appropriate saddle-point of Ψ, which itself takes the form:

$$\Psi[\ldots] = i\sum_\alpha \hat{Q}_\alpha(1-q_{\alpha\alpha}) + iR\sum_\alpha \hat{R}_\alpha + i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\,q_{\alpha\beta} + i\sum_\alpha \int dxdy\ \hat{P}_\alpha(x,y)P[x,y]$$

$$+\alpha\log\mathcal{D}[0,0] + \lim_{N\to\infty}\frac{1}{N}\sum_i \log\int d\boldsymbol{\sigma}\ e^{-i\tau_i\sqrt{Q}\sum_\alpha \hat{R}_\alpha\sigma_\alpha - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta} \quad (4.11)$$

With $\mathcal{D}[u,v]$ given by (4.6) and with the function $\mathcal{L}[u,v;u',v']$ given by (4.5). The auxiliary order parameters $q_{\alpha\beta}$ have the usual interpretation in terms of the average probability density for finding a mutual overlap q of two independently evolving weight vectors with the same realization of the training set (see e.g. Mézard et al, 1987):

$$\langle P(q)\rangle_{\Xi} = \left\langle \left\langle\!\!\left\langle \delta\left[q - \frac{\boldsymbol{J}^a\cdot\boldsymbol{J}^b}{|\boldsymbol{J}^a||\boldsymbol{J}^b|}\right]\right\rangle\!\!\right\rangle\right\rangle_{\Xi} = \lim_{n\to 0}\frac{1}{n(n-1)}\sum_{\alpha\neq\beta}\delta[q-q_{\alpha\beta}] \quad (4.12)$$

We now make the replica symmetric (RS) Ansatz in the extremization problem, which according to (4.12) is equivalent to assuming ergodicity. With a modest amount of foresight we put

$$q_{\alpha\beta} = q_0\delta_{\alpha\beta} + q[1-\delta_{\alpha\beta}] \qquad \hat{q}_{\alpha\beta} = \frac{i}{2}[r - r_0\delta_{\alpha\beta}]$$

$$\hat{R}_\alpha = i\rho \qquad \hat{Q}_\alpha = i\phi \qquad \hat{P}_\alpha(u,v) = i\chi[u,v]$$

This allows us to expand the quantity Ψ of (4.11) for small n:

$$\lim_{n\to 0}\frac{1}{n}\Psi[\ldots] = -\phi(1-q_0) - \rho R + \frac{1}{2}qr - \frac{1}{2}q_0(r-r_0) - \frac{1}{2}\log r_0 + \frac{1}{2r_0}(r+\rho^2 Q)$$

$$-\int dxdy\ \chi[x,y]\ P[x,y] + \lim_{n\to 0}\frac{\alpha}{n}\log\mathcal{D}[0,0] + \text{constants}$$

At this stage it is useful to work out those saddle-point equations that follow upon variation of $\{\phi, r, \rho, r_0\}$:

$$q_0 = 1 \qquad r_0 = \frac{1}{1-q} \qquad \rho = \frac{R}{Q(1-q)} \qquad r = \frac{qQ-R^2}{Q(1-q)^2}$$

These allow us to eliminate most variational parameters, leaving a saddle-point problem involving only the function $\chi[x,y]$ and the scalar q:

$$\lim_{n\to 0}\ \frac{1}{n}\Psi[q,\{\chi\}] = \frac{1-R^2/Q}{2(1-q)} + \frac{1}{2}\log(1-q) - \int dxdy\ \chi[x,y]\ P[x,y]$$

$$+ \lim_{n\to 0}\ \frac{\alpha}{n}\log \mathcal{D}[0,0;q,\{\chi\}] + \text{constants} \tag{4.13}$$

Finally we have to work out the RS version of $\mathcal{D}[0,0;q,\{\chi\}]$, as defined more generally in (4.6). The inverse of the matrix in (4.8), in RS Ansatz, is found to be:

$$\boldsymbol{A} = \begin{pmatrix} C_{11} & \cdots & C_{1n} & \gamma \\ \vdots & & \vdots & \vdots \\ C_{n1} & \cdots & C_{nn} & \gamma \\ \gamma & \cdots & \gamma & b \end{pmatrix} \qquad C_{\alpha\beta} = \frac{\delta_{\alpha\beta}}{1-q} - d \qquad \begin{array}{l} \gamma = -\frac{R/\sqrt{Q}}{1-q} + \mathcal{O}(n) \\ b = 1 + \mathcal{O}(n) \\ d = \frac{q-R^2/Q}{(1-q)^2} + \mathcal{O}(n) \end{array} \tag{4.14}$$

With this expression we obtain

$$\mathcal{D}[0,0;q,\{\chi\}] = \frac{\int d\boldsymbol{x}\ dy\ e^{-\frac{1}{2}\boldsymbol{x}\cdot\boldsymbol{C}\boldsymbol{x}-\frac{1}{2}by^2-\gamma y\sum_\alpha x_\alpha+\frac{1}{\alpha}\sum_\alpha \chi(\sqrt{Q}x_\alpha,y)}}{\int d\boldsymbol{x}\ dy\ e^{-\frac{1}{2}\boldsymbol{x}\cdot\boldsymbol{C}\boldsymbol{x}-\frac{1}{2}by^2-\gamma y\sum_\alpha x_\alpha}}$$

$$= \frac{\int DzDy\left[\int dx\ e^{-\frac{x^2}{2(1-q)}+[z\sqrt{d}-\gamma\frac{y}{\sqrt{b}}]x+\frac{1}{\alpha}\chi(\sqrt{Q}x,\frac{y}{\sqrt{b}})}\right]^n}{\int DzDy\left[\int dx\ e^{-\frac{1}{2(1-q)}x^2+[z\sqrt{d}-\gamma\frac{y}{\sqrt{b}}]x}\right]^n}$$

$$\lim_{n\to 0}\frac{\alpha}{n}\log\mathcal{D}[0,0;q,\{\chi\}] = \alpha\int DzDy\log\left\{\frac{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[z\sqrt{d}-\gamma y]/\sqrt{Q}+\frac{1}{\alpha}\chi(x,y)}}{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[z\sqrt{d}-\gamma y]/\sqrt{Q}}}\right\}$$

with the usual short-hand $Dy = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}y^2}$. We can simplify this result by defining

$$A = R/Q(1-q) \qquad\qquad B = \sqrt{qQ-R^2}/Q(1-q) \tag{4.15}$$

which gives

$$\lim_{n\to 0}\frac{\alpha}{n}\log\mathcal{D}[0,0;q,\{\chi\}] = \alpha\int DzDy\ \log\left\{\frac{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi(x,y)}}{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]}}\right\}$$

Upon carrying out the x-integration in the denominator of this expression we can write (4.13) in a surprisingly simple form (with the short-hands (4.15)):

$$\lim_{n\to 0}\frac{1}{n}\Psi[q,\{\chi\}] = \frac{1-\alpha-R^2/Q}{2(1-q)} + \frac{1}{2}(1-\alpha)\log(1-q) - \int dxdy\ \chi[x,y]\ P[x,y]$$

$$+\ \alpha\int DzDy\ \log\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]} \tag{4.16}$$

Note that (4.16) is to be *minimized*, both with respect to q (which originated as an $n(n-1)$-fold entry in a matrix, leading to curvature sign change for $n<1$) and with respect to $\chi[x,y]$ (obtained from the n-fold occurrence of the function $\hat{P}[x,y]$, multiplied by i, which also leads to a curvature sign change).

The remaining saddle point equations are obtained by (functional) variation with respect to χ:

$$\text{for all } x,y: \qquad P[x,y]=\frac{e^{-\frac{1}{2}y^2}}{\sqrt{2\pi}}\int Dz\left\{\frac{e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}{\int dx'\ e^{-\frac{x'^2}{2Q(1-q)}+x'[Ay+Bz]+\frac{1}{\alpha}\chi[x',y]}}\right\}, \tag{4.17}$$

and q (using equation (4.17) wherever possible):

$$\int dxdy\ P[x,y](x-Ry)^2+(R^2-qQ)(\frac{1}{\alpha}-1)=$$

$$\left[2\sqrt{qQ-R^2}+\frac{Q(1-q)}{\sqrt{qQ-R^2}}\right]\int DyDz\ z\left\{\frac{\int dx\ x\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}\right\} \tag{4.18}$$

Apart from the physically irrelevant degree of freedom $\chi[x,y]\to\chi[x,y]+\rho(y)$, for arbitrary $\rho(y)$, the solution of the functional saddle-point problem (4.17), if it exists, will be unique for any given value of q in the physical range $R^2/Q\leq q\leq 1$. This follows immediately from the convexity of $\Psi[\ldots]$ (4.16), which can, in turn, be deduced from the the fact that the second functional derivative of $\Psi[\ldots]$ with respect to the function $\chi[\ldots]$ is a non-negative operator. In addition one can prove that $\Psi[\ldots]$ (4.16) has a lower bound, which is given in terms of the differential entropy of the distribution $P[x,y]$. Furthermore, the functional saddle-point equation (4.17) can be rewritten in the form of a fixed-point equation associated with an iterative mapping for the function $\chi[x,y]$, such that this mapping has (4.16) as a Lyapunov functional. Should an analytical solution of (4.17) turn out to be impossible, in combination the above properties convert finding the solution of (4.17) from a potentially insurmountable obstacle into a straightforward numerical exercise. Details will be published in (Coolen and Saad, 1998).

4.3 Explicit Expression for the Green's Function

In order to work out the Green's function (4.9) we need $\mathcal{L}[u,v;u',v']$ as defined in (4.5) which, in turn, is given in terms of the integrals (4.6,4.7). First we calculate in RS ansatz the $n\to 0$ limit of $D[u,v;q,\{\chi\}]$ (4.6), using (4.14), and simplify the result with the saddle-point equation (4.17):

$$\lim_{n\to 0}\mathcal{D}[u,v;q,\{\chi\}]=\int DzDy\ e^{-ivy}\ \frac{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]-iux}}{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}$$

$$= \int dxdy\ P[x,y]\ e^{-ivy-iux} \tag{4.19}$$

Next we work out $F_\lambda^\alpha[u,v]$ (4.7) in RS Ansatz, using (4.14), with $\lambda \in \{1,2\}$, which results in

$$\lim_{n\to 0} \mathcal{F}_\lambda^\alpha[u,v] = i \lim_{n\to 0}$$
$$\int DyDz\ e^{-ivy} \int d\boldsymbol{x}\ e^{\sum_\beta \left[-\frac{1}{2}\frac{x_\beta^2}{1-q}+[z\sqrt{d}-\gamma y]x_\beta+\frac{1}{\alpha}\chi[\sqrt{Q}x_\beta,y]\right]-iux_1\sqrt{Q}} \partial_\lambda \chi[\sqrt{Q}x_\alpha, y]$$

Replica permutation symmetries allow us to simplify this expression:

$$\lim_{n\to 0} \mathcal{F}_\lambda^\alpha[u,v] = \delta_{\alpha 1} F_\lambda^1[u,v] + (1-\delta_{\alpha 1}) F_\lambda^2[u,v] \tag{4.20}$$

with

$$F_\lambda^1[u,v] = i \int dx\ dy\ P[x,y]\ e^{-ivy-iux}\ \partial_\lambda \chi[x,y] \tag{4.21}$$

and

$$F_\lambda^2[u,v] = i \int DyDz\ e^{-ivy}$$
$$\frac{\left[\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}\ \partial_\lambda \chi[x,y]\right]\left[\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]-iux}\right]}{\left[\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}\right]^2} \tag{4.22}$$

We can now proceed with the calculation of (4.5), whose building blocks are

$$\alpha^{-1}\mathcal{F}_1^\alpha[u,v] + u\delta_{\alpha 1}\mathcal{D}[u,v] = \delta_{\alpha 1} G_1[u,v] + (1-\delta_{\alpha 1})\tilde{G}_{1,2}[u,v]$$
$$\alpha^{-1}\mathcal{F}_2^\alpha[u,v] + v\delta_{\alpha 1}\mathcal{D}[u,v] = \delta_{\alpha 1} G_2[u,v] + (1-\delta_{\alpha 1})\tilde{G}_2[u,v]$$

with

$$G_1[u,v] = \alpha^{-1}\mathcal{F}_{1,2}^1[u,v] + u\mathcal{D}[u,v] \qquad \tilde{G}_1[u,v] = \alpha^{-1}\mathcal{F}_1^2[u,v]$$
$$G_2[u,v] = \alpha^{-1}\mathcal{F}_2^1[u,v] + v\mathcal{D}[u,v] \qquad \tilde{G}_2[u,v] = \alpha^{-1}\mathcal{F}_2^2[u,v]$$

and their Fourier transforms:

$$\hat{G}_1[\hat{u},\hat{v}] = \int \frac{du\ dv}{(2\pi)^2} e^{iu\hat{u}+iv\hat{v}} G_1[u,v] \qquad \bar{G}_1[\hat{u},\hat{v}] = \int \frac{du\ dv}{(2\pi)^2} e^{iu\hat{u}+iv\hat{v}} \tilde{G}_1[u,v]$$
$$\hat{G}_2[\hat{u},\hat{v}] = \int \frac{du\ dv}{(2\pi)^2} e^{iu\hat{u}+iv\hat{v}} G_2[u,v] \qquad \bar{G}_2[\hat{u},\hat{v}] = \int \frac{du\ dv}{(2\pi)^2} e^{iu\hat{u}+iv\hat{v}} \tilde{G}_2[u,v]$$

With these short-hands we obtain a relatively compact expression for (4.5). In this expression we can subsequently take the limit $n \to 0$, insert the result into our equation (4.9) for the Green's function $\mathcal{A}[x,y;x',y']$, and find:

$$\mathcal{A}[x,y;x',y'] = -Q(1-q)\left[\hat{G}_1[x,y]\hat{G}_1[x',y'] - \bar{G}_1[x,y]\bar{G}_1[x',y']\right]$$
$$-Qq\left[\hat{G}_1[x,y] - \bar{G}_1[x,y]\right]\left[\hat{G}_1[x',y'] - \bar{G}_1[x',y']\right]$$
$$-R\left[\hat{G}_1[x,y] - \bar{G}_1[x,y]\right]\left[\hat{G}_2[x',y'] - \bar{G}_2[x',y']\right]$$
$$-R\left[\hat{G}_1[x',y'] - \bar{G}_1[x',y']\right]\left[\hat{G}_2[x,y] - \bar{G}_2[x,y]\right]$$
$$-\left[\hat{G}_2[x,y] - \bar{G}_2[x,y]\right]\left[\hat{G}_2[x',y'] - \bar{G}_2[x',y']\right] \qquad (4.23)$$

Finally, working out the four relevant Fourier transforms, using (4.19,4.21,4.22), gives:

$$\hat{G}_1[x,y] = i\left[\frac{1}{\alpha}\ P[x,y]\ \frac{\partial}{\partial x}\chi[x,y] - \frac{\partial}{\partial x}P[x,y]\right]$$
$$\hat{G}_2[x,y] = i\left[\frac{1}{\alpha}\ P[x,y]\ \frac{\partial}{\partial y}\chi[x,y] - \frac{\partial}{\partial y}P[x,y]\right]$$

$$\bar{G}_1[x,y] = \frac{i}{\alpha}\frac{e^{-\frac{1}{2}y^2}}{\sqrt{2\pi}}\int Dz$$
$$\frac{\left[\int dx'\ e^{-\frac{x'^2}{2Q(1-q)}+x'[Ay+Bz]+\frac{1}{\alpha}\chi[x',y]}\ \partial_1\chi[x',y]\right]e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}{\left[\int dx'\ e^{-\frac{x'^2}{2Q(1-q)}+x'[Ay+Bz]+\frac{1}{\alpha}\chi[x',y]}\right]^2}$$

$$\bar{G}_2[x,y] = \frac{i}{\alpha}\frac{e^{-\frac{1}{2}y^2}}{\sqrt{2\pi}}\int Dz$$
$$\frac{\left[\int dx'\ e^{-\frac{x'^2}{2Q(1-q)}+x'[Ay+Bz]+\frac{1}{\alpha}\chi[x',y]}\ \partial_2\chi[x',y]\right]e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}{\left[\int dx'\ e^{-\frac{x'^2}{2Q(1-q)}+x'[Ay+Bz]+\frac{1}{\alpha}\chi[x',y]}\right]^2}$$

4.4 Simplification and Summary of the Theory

In this section we simplify and summarize the results obtained so far. Since the distribution $P[x,y]$ obeys $P[x,y] = P[x|y]P[y]$ with $P[y] = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}y^2}$, our equations can be simplified by choosing as our order parameter function the conditional distribution $P[x|y]$. We also replace the conjugate order parameter function $\chi[x,y]$ by the effective measure $M[x,y]$, and we introduce a compact notation for the relevant averages in our problem:

$$M[x,y] = e^{-\frac{x^2}{2Q(1-q)}+Axy+\frac{1}{\alpha}\chi[x,y]} \qquad \langle f[x,y,z]\rangle_\star = \frac{\int dx\ M[x,y]e^{Bxz}f[x,y,z]}{\int dx\ M[x,y]e^{Bxz}}$$

Instead of the original Green's function $\mathcal{A}[x,y;x',y']$ we turn to the transformed Green's function $\tilde{\mathcal{A}}[x,y;x',y']$, defined as

$$\mathcal{A}[x,y;x',y'] = P[x,y]\tilde{\mathcal{A}}[x,y;x',y']P[x',y']$$

With these notational conventions one finds that (4.23) translates into

$$\begin{aligned}\tilde{\mathcal{A}}[x,y;x',y'] &= Q(1-q)\left[J_1[x,y]J_1[x',y']-\tilde{J}_1[x,y]\tilde{J}_1[x',y']\right]+J_2[x,y]J_2[x',y']\\ &+R\left[J_1[x,y]-\tilde{J}_1[x,y]\right]J_2[x',y']+R\left[J_1[x',y']-\tilde{J}_1[x',y']\right]J_2[x,y]\\ &+Qq\left[J_1[x,y]-\tilde{J}_1[x,y]\right]\left[J_1[x',y']-\tilde{J}_1[x',y']\right]\end{aligned} \tag{4.24}$$

with

$$J_1[x,y] = \frac{\partial}{\partial x}\log\frac{M[x,y]}{P[x|y]}+\frac{x-Ry}{Q(1-q)}$$

$$\tilde{J}_1[X,y] = P[X|y]^{-1}\int Dz\ \langle\frac{\partial}{\partial x}\log M[x,y]+\frac{x-Ry}{Q(1-q)}\rangle_\star\langle\delta[X-x]\rangle_\star$$

$$\begin{aligned}J_2[X,y] &= \frac{\partial}{\partial y}\log\frac{M[X,y]}{P[X|y]}-\frac{RX}{Q(1-q)}+y\\ &-P[X|y]^{-1}\int Dz\ \langle\frac{\partial}{\partial y}\log M[x,y]-\frac{Rx}{Q(1-q)}\rangle_\star\langle\delta[X-x]\rangle_\star\end{aligned}$$

It turns out that significant simplification of the result (4.24) is possible, upon using the following two identities:

$$\langle\frac{\partial}{\partial x}\log M[x,y]\rangle_\star = -Bz$$

$$\langle\frac{\partial}{\partial y}\log M[x,y]\rangle_\star = \frac{\partial}{\partial y}\log\int dx\ e^{Bxz}M[x,y]$$

To achieve the desired simplification of $\tilde{\mathcal{A}}[x,y;x',y']$ we define

$$\Phi[X,y] = \left\{Q(1-q)P[X|y]\right\}^{-1}\int Dz\ \langle X-x\rangle_\star\langle\delta[X-x]\rangle_\star \tag{4.25}$$

We can now, after additional integration by parts with respect to z, simplify the above expressions for $J_1[\ldots]$, $\tilde{J}_1[\ldots]$ and $J_2[\ldots]$ to

$$J_1[x,y] = \frac{x-Ry}{Q(1-q)}-\frac{qQ-R^2}{Q(1-q)}\Phi[x,y] \qquad \tilde{J}_1[x,y] = J_1[x,y]-\Phi[x,y]$$

$$J_2[x,y] = y - R\Phi[x,y]$$

and consequently

$$\tilde{A}[x,y;x',y'] = yy'+(x-Ry)\Phi[x',y']+(x'-Ry')\Phi[x,y]-(Q-R^2)\Phi[x,y]\Phi[x',y'] \tag{4.26}$$

The kernel $\tilde{A}[x,y;x',y']$ is now written in an explicitly separable form, as a result of which our theory can be summarized in just a single page:

• Philosophy and Notation

Our observables $Q=\boldsymbol{J}^2$, $R=\boldsymbol{B}\cdot\boldsymbol{J}$ and $P[x,y]=\langle\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}}$ obey deterministic and self-averaging laws for $N\to\infty$, with $P[y]=(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}y^2}$. We abbreviate $\langle f[x,y]\rangle=\int dxDy\ P[x|y]f[x,y]$ and (with Φ defined below):

$$U=\langle\Phi[x,y]\mathcal{G}[x,y]\rangle \qquad V=\langle x\mathcal{G}[x,y]\rangle \qquad W=\langle y\mathcal{G}[x,y]\rangle \qquad Z=\langle\mathcal{G}^2[x,y]\rangle$$

The training- and generalization errors are given by

$$E_{\rm t}=\langle\theta[-xy]\rangle \qquad\qquad E_{\rm g}=\pi^{-1}\arccos[R/\sqrt{Q}] \tag{4.27}$$

• Macroscopic Dynamic Equations

On-line learning:

$$\frac{d}{dt}Q=2\eta V+\eta^2 Z \qquad\qquad \frac{d}{dt}R=\eta W \tag{4.28}$$

$$\frac{d}{dt}P[x|y]=\frac{1}{\alpha}\int dx'P[x'|y]\left\{\delta[x-x'-\eta\mathcal{G}[x',y]]-\delta[x-x']\right\}+\frac{1}{2}\eta^2 Z\frac{\partial^2}{\partial x^2}P[x|y]$$
$$-\eta\frac{\partial}{\partial x}\left\{P[x|y]\left[U(x-Ry)+Wy\right]\right\}-\eta\left[V-RW-(Q-R^2)U\right]\frac{\partial}{\partial x}\left\{P[x|y]\Phi[x,y]\right\} \tag{4.29}$$

Batch learning:

$$\frac{d}{dt}Q=2\eta V \qquad\qquad \frac{d}{dt}R=\eta W \tag{4.30}$$

$$\frac{d}{dt}P[x|y]=-\frac{\eta}{\alpha}\frac{\partial}{\partial x}\left\{P[x|y]\mathcal{G}[x,y]\right\}-\eta\frac{\partial}{\partial x}\left\{P[x|y]\left[U(x-Ry)+Wy\right]\right\}$$
$$-\eta\left[V-RW-(Q-R^2)U\right]\frac{\partial}{\partial x}\left\{P[x|y]\Phi[x,y]\right\} \tag{4.31}$$

• Saddle-Point Equations and the Function Φ

The key function $\Phi[x,y]$ occurring in the above equations is given by

$$\Phi[X,y]=\left\{Q(1-q)P[X|y]\right\}^{-1}\int Dz\ \langle X-x\rangle_\star\langle\delta[X-x]\rangle_\star \tag{4.32}$$

with

$$\langle f[x,y,z]\rangle_\star=\frac{\int dx\ M[x,y]e^{Bxz}f[x,y,z]}{\int dx\ M[x,y]e^{Bxz}} \qquad B=\frac{\sqrt{qQ-R^2}}{Q(1-q)} \tag{4.33}$$

The spin-glass order parameter $q\in[R^2/Q,1]$ and the function $M[x,y]$ are calculated at each time-step by solving the saddle-point equations

$$\langle(x-Ry)^2\rangle+(qQ-R^2)(1-\frac{1}{\alpha})=\left[2(qQ-R^2)^{\frac{1}{2}}+\frac{1}{B}\right]\int DyDz\ z\langle x\rangle_\star \tag{4.34}$$

$$P[X|y]=\int Dz\ \langle\delta[X-x]\rangle_\star \tag{4.35}$$

5 Tests and Applications of the Theory

5.1 Locally Gaussian Solutions

There are two advantages of rewriting our equations in Fourier representation. Firstly, the functional saddle-point equation (4.35) will acquire a simpler form. Secondly, in those cases where we expect $P[x|y]$ to be of a Gaussian form in x this will simplify solution of the diffusion equations (4.29,4.31). Clearly, $P[x,y]$ being Gaussian in (x,y) is not equivalent to $P[x|y]$ being Gaussian in x only. The former will only turn out to occur for $\alpha\to\infty$. A Gaussian $P[x|y]$ with moments which depend in a non-trivial way on y, on the other hand, can also occur for $\alpha<\infty$, provided we consider simple learning rules and small η. To avoid ambiguity we will call solutions of the latter type 'locally Gaussian'.

We normalize the measure $M[x,y]$ such that $\int dx\ M[x,y]=1$ for all $y\in\Re$, emphasizing the result in our notation by writing $M[x,y]\to M[x|y]$, and we introduce the Fourier transforms

$$\hat{P}[k|y]=\int dx\ e^{-ikx}P[x|y] \qquad\qquad \hat{M}[k|y]=\int dx\ e^{-ikx}M[x|y]$$

The transformed functional saddle-point equation thereby becomes

$$\hat{P}[k|y]=\int Dz\ \frac{\hat{M}[k+iBz|y]}{\hat{M}[iBz|y]} \tag{5.1}$$

Transformation of the on-line equation (4.29) for $P[x|y]$ (from the which the batch equation (4.31) can be obtained by expansion in η) gives:

$$\frac{d}{dt}\log\hat{P}[k|y]\ =\ \frac{1}{\alpha}\left\{\int dk'\ \frac{\hat{P}[k'|y]}{\hat{P}[k|y]}\int\frac{dx'}{2\pi}e^{ix'(k'-k)-i\eta k\mathcal{G}[x',y]}-1\right\}-i\eta k(W\!-\!UR)y$$

$$-\frac{1}{2}\eta^2k^2Z+\eta kU\frac{\partial}{\partial k}\log\hat{P}[k|y]-i\eta k\left[\frac{V-RW-(Q-R^2)U}{\sqrt{qQ-R^2}\hat{P}[k|y]}\right]\int Dz\ z\ \frac{\hat{M}[k+iBz|y]}{\hat{M}[iBz]} \tag{5.2}$$

If $P[x|y]$ is Gaussian in x we can solve the functional saddle-point equation (4.35) (whose solution is unique), and find

$$P[x|y]=\frac{e^{-\frac{1}{2}[x-\overline{x}(y)]^2/\Delta^2(y)}}{\Delta(y)\sqrt{2\pi}} \qquad\qquad M[x|y]=\frac{e^{-\frac{1}{2}[x-\overline{x}(y)]^2/\sigma^2(y)}}{\sigma(y)\sqrt{2\pi}} \tag{5.3}$$

$$\Delta^2(y)=\sigma^2(y)+B^2\sigma^4(y) \tag{5.4}$$

with $\hat{P}[k|y]=\exp\left[-ik\overline{x}(y)-\frac{1}{2}k^2\Delta^2(y)\right]$ and $\hat{M}[k|y]=\exp\left[-ik\overline{x}(y)-\frac{1}{2}k^2\sigma^2(y)\right]$. Insertion of these expression as an Ansatz into (5.2), using the identity

$$\int Dz\ z\ \frac{\hat{M}[k+iBz|y]}{\hat{M}[iBz]}=ikB\sigma^2(y)\hat{P}[k|y]$$

(which holds only for locally Gaussian solutions) and performing some simple

manipulations, gives the simplified equation

$$-ik\frac{d}{dt}\overline{x}(y)-\frac{1}{2}k^2\frac{d}{dt}\Delta^2(y) \;=\; \frac{1}{\alpha}\left\{\int\frac{du}{\sqrt{2\pi}}e^{-\frac{1}{2}[u-ik\Delta(y)]^2-ik\eta\mathcal{G}[\overline{x}(y)+u\Delta(y),y]}-1\right\}$$
$$-i\eta k\left\{Wy+U[\overline{x}(y)-Ry]\right\}$$
$$-\frac{1}{2}k^2\left\{\eta^2 Z+2\eta U\Delta^2(y)+2\eta\sigma^2(y)\left[\frac{V-RW-(Q-R^2)U}{Q(1-q)}\right]\right\} \qquad (5.5)$$

It follows that locally Gaussian solutions can occur in two situations only:

$$\alpha=\infty \qquad \text{or} \qquad \frac{\partial^3}{\partial k^3}\int\frac{du}{\sqrt{2\pi}}\;e^{-\frac{1}{2}[u-ik\Delta(y)]^2-ik\eta\mathcal{G}[\overline{x}(y)+u\Delta(y),y]}=0$$

The first case corresponds to complete training sets (see next section). The second case occurs for sufficiently simple learning rules $\mathcal{G}[x,y]$, in combination either with batch execution (so that we retain only the term linear in η) or with on-line execution for small η (retaining only η and η^2 terms).

5.2 Link with the Complete Training Sets Formalism

The least we should require of our theory is that it reduces to the simple formalism of complete training sets in the limit $\alpha\to\infty$. In the previous section we have seen that for $\alpha\to\infty$ our driven diffusion equations for the conditional distribution $P[x|y]$ have locally Gaussian solutions, with $\int dx\; xP[x|y]=\overline{x}(y)$ and $\int dx[x-\overline{x}(y)]^2P[x|y]=\Delta^2(y)$. Note that for such solutions we can calculate objects such as $\langle x\rangle_\star$ and the function $\Phi[x,y]$ of (4.32) directly, giving

$$\langle x\rangle_\star=\overline{x}(y)+zB\sigma^2(y) \qquad\qquad \Phi[x,y]=\frac{x-\overline{x}(y)}{Q(1-q)[1+B^2\sigma^2(y)]}$$

with $\Delta^2(y)=\sigma^2(y)+B^2\sigma^4(y)$ and $B=\sqrt{qQ-R^2}/Q(1-q)$. The remaining equations to be solved are those for Q and R, in combination with dynamical equations for the y-dependent cumulants $\overline{x}(y)$ and $\Delta^2(y)$. These reduce to:

$$\frac{d}{dt}Q=\begin{cases}2\eta\langle x\mathcal{G}[x,y]\rangle+\eta^2\langle\mathcal{G}^2[x,y]\rangle & (\text{on-line})\\ 2\eta\langle x\mathcal{G}[x,y]\rangle & (\text{batch})\end{cases} \qquad \frac{d}{dt}R=\eta\langle y\mathcal{G}[x,y]\rangle \qquad (5.6)$$

$$\frac{1}{\eta}\frac{d}{dt}\Big[\overline{x}(y)-Ry\Big]=[\overline{x}(y)-Ry]\langle\Phi[x',y']\mathcal{G}[x',y']\rangle \qquad (5.7)$$

$$\frac{1}{2\eta}\frac{d}{dt}\Big[\Delta^2(y)-Q+R^2\Big]=\langle(x'-Ry')\mathcal{G}[x',y']\rangle\left[\frac{\sigma^2(y)}{Q(1-q)}-1\right]$$
$$+\langle\Phi[x',y']\mathcal{G}[x',y']\rangle\left[\Delta^2(y)-\frac{Q-R^2}{Q(1-q)\sigma^2(y)}\right] \qquad (5.8)$$

with one remaining saddle-point equation to determine q, obtained upon working out (4.34) for locally Gaussian solutions:

$$\int Dy\,\Big\{[\overline{x}(y)-Ry]^2+\Delta^2(y)\Big\}+qQ-R^2=\left[2\frac{qQ-R^2}{Q(1-q)}+1\right]\int Dy\;\sigma^2(y) \quad (5.9)$$

We now make the Ansatz that $\overline{x}(y)=Ry$ and $\Delta^2(y)=Q-R^2$, i.e.

$$P[x|y]=\frac{e^{-\frac{1}{2}[x-Ry]^2/(Q-R^2)}}{\sqrt{2\pi(Q-R^2)}}, \quad (5.10)$$

Insertion into the dynamical equations shows that (5.7) is now immediately satisfied, that (5.8) reduces to $\sigma^2(y)=Q(1-q)$, and that the saddle-point equation (5.9) is automatically satisfied. Since (5.10) is parametrized by Q and R only, the equations (5.6) are closed. From our general theory for restricted training sets we thus indeed recover in the limit $\alpha\to\infty$ the standard formalism (5.6,5.10) describing learning with complete training sets, as claimed.

5.3 Benchmark Tests: Hebbian Learning

In the special case of the Hebb rule, $\mathcal{G}[x,y]=\mathrm{sgn}[y]$, where weight changes $\Delta\boldsymbol{J}$ never depend on $\boldsymbol{J}$, one can write down an explicit expression for the weight vector $\boldsymbol{J}$ at any time, and thus for the expectation values of our observables. We choose as our initial field distribution a simple Gaussian one, resulting from an initialization process which did not involve the training set:

$$P_0[x|y]=\frac{e^{-\frac{1}{2}(x-R_0y)^2/(Q_0-R_0^2)}}{\sqrt{2\pi(Q_0-R_0^2)}} \quad (5.11)$$

Careful averaging of the exact expressions for our observables over all 'paths' $\{\boldsymbol{\xi}(0),\boldsymbol{\xi}(1),\ldots\}$ taken by the question vector through the training set $\tilde{D}$ (for on-line learning), followed by averaging over all realizations of the training set $\tilde{D}$ of size $p=\alpha N$, and taking the $N\to\infty$ limit, then leads to the following *exact* result (Rae et al, 1998). For on-line Hebbian learning one ends up with:

$$Q=Q_0+2\eta tR_0\sqrt{\frac{2}{\pi}}+\eta^2t+\eta^2t^2\left[\frac{1}{\alpha}+\frac{2}{\pi}\right] \qquad R=R_0+\eta t\sqrt{\frac{2}{\pi}} \quad (5.12)$$

$$P[x|y]=\int\frac{d\hat{x}}{2\pi}\;e^{-\frac{1}{2}\hat{x}^2[Q-R^2]+i\hat{x}[x-Ry]+\frac{t}{\alpha}[e^{-i\eta\hat{x}\;\mathrm{sgn}[y]}-1]} \quad (5.13)$$

For batch learning a similar calculation [3] gives:

$$Q=Q_0+2\eta tR_0\sqrt{\frac{2}{\pi}}+\eta^2t^2\left[\frac{1}{\alpha}+\frac{2}{\pi}\right] \qquad R=R_0+\eta t\sqrt{\frac{2}{\pi}} \quad (5.14)$$

[3] Note that in Rae et al (1998) only the on-line calculation was carried out; the batch calculation can be done along the same lines.

$$P[x|y] = \frac{e^{-\frac{1}{2}[x-Ry-(\eta t/\alpha)\ \mathrm{sgn}[y]]^2/(Q-R^2)}}{\sqrt{2\pi(Q-R^2)}} \tag{5.15}$$

Neither of the two field distributions is of a fully Gaussian form (although the batch distribution is at least locally Gaussian). Note that for both on-line and batch Hebbian learning we have

$$\int dx\ xP[x|y] = Ry + \frac{\eta t}{\alpha}\ \mathrm{sgn}[y] \tag{5.16}$$

The generalization- and training errors are, as before, given in terms of the above observables as $E_{\rm g} = \pi^{-1}\arccos[R/\sqrt{Q}]$ and $E_{\rm t} = \int Dydx P[x|y]\theta[-xy]$. We thus have exact expressions for both the generalization error and the training error at any time and for any α. Their asymptotic values are, for both batch and on-line Hebbian learning, given by

$$\lim_{t\to\infty} E_{\rm g} = \frac{1}{\pi}\arccos\left[\frac{1}{\sqrt{1+\pi/2\alpha}}\right] \tag{5.17}$$

$$\lim_{t\to\infty} E_{\rm t} = \frac{1}{2} - \frac{1}{2}\int Dy\ \mathrm{erf}\left[|y|\sqrt{\frac{\alpha}{\pi}}+\frac{1}{\sqrt{2\alpha}}\right] \tag{5.18}$$

As far as $E_{\rm g}$ and $E_{\rm t}$ are concerned, the differences between batch and on-line Hebbian learning are confined to transients. Clearly, the above exact results (which can only be obtained for Hebbian-type learning rules) provide excellent and welcome benchmarks with which to test general theories such as ours.

5.4 Batch Hebbian Learning

We now compare the exact solutions for Hebbian learning to the predictions of our general theory, turning first to batch Hebbian learning. We insert into the equations of our general formalism the Hebbian recipe $\mathcal{G}[x,y] = \mathrm{sgn}[y]$. This simplifies our dynamic equations enormously. In particular we obtain:

$$U = 0, \qquad V = \langle x\ \mathrm{sgn}(y)\rangle, \qquad W = \sqrt{2/\pi}$$

For batch learning we consequently find:

$$\frac{d}{dt}Q = 2\eta V \qquad\qquad \frac{d}{dt}R = \eta\sqrt{2/\pi}$$

$$\frac{d}{dt}P[x|y] = -\frac{\eta}{\alpha}\ \mathrm{sgn}(y)\frac{\partial}{\partial x}P[x|y] - \eta y\sqrt{\frac{2}{\pi}}\frac{\partial}{\partial x}P[x|y]$$

$$-\eta(V-R\sqrt{\frac{2}{\pi}})\frac{\partial}{\partial x}\left\{P[x|y]\Phi[x,y]\right\}$$

Given the initial field distribution (5.11), we immediate derive $V_0 = R_0\sqrt{2/\pi}$. From the general property $\int dx\ P[x|y]\Phi[x,y] = 0$ and the above diffusion equation for $P[x|y]$ we derive an equation for the quantity $V = \langle x\ \mathrm{sgn}(y)\rangle$,

resulting in $\frac{d}{dt}V = \eta/\alpha + 2\eta/\pi$, which subsequently allows us to solve

$$Q = Q_0 + 2\eta t R_0\sqrt{\frac{2}{\pi}} + \eta^2 t^2\left[\frac{1}{\alpha}+\frac{2}{\pi}\right] \qquad R = R_0 + \eta t\sqrt{\frac{2}{\pi}} \tag{5.19}$$

Furthermore, it turns out that the above diffusion equation for $P[x|y]$ obeys the conditions for having locally Gaussian solutions, i.e.

$$P[x|y] = \frac{e^{-\frac{1}{2}[x-\overline{x}(y)]^2/\Delta^2(y)}}{\Delta(y)\sqrt{2\pi}}, \qquad M[x|y] = \frac{e^{-\frac{1}{2}[x-\overline{x}(y)]^2/\sigma^2(y)}}{\sigma(y)\sqrt{2\pi}}$$

provided the y-dependent average $\overline{x}(y)$ and the y-dependent variances $\Delta(y)$ and $\sigma(y)$ obey the following three equations:

$$\overline{x}(y) = Ry + \frac{\eta t}{\alpha}\,\mathrm{sgn}(y) \qquad \frac{d}{dt}\Delta^2(y) = \frac{2\eta^2 t\sigma^2(y)}{\alpha Q(1-q)}$$

$$\Delta^2(y) = \sigma^2(y) + B^2\sigma^4(y)$$

The spin-glass order parameter q is to be solved from the remaining saddle-point equation. With help of identities like $\langle x\rangle_\star = \overline{x}(y) + zB\sigma^2(y)$, which only hold for locally Gaussian solutions, one can simplify the latter to

$$\frac{\eta^2 t^2}{\alpha} + \alpha\int Dy\;\Delta^2(y) + (qQ-R^2)(\alpha-1) = \alpha\left[2\frac{qQ-R^2}{Q(1-q)}+1\right]\int Dy\;\sigma^2(y)$$

We now immediately find the solution

$$\Delta^2(y) = Q-R^2, \qquad \sigma^2(y) = Q(1-q), \qquad q = [\alpha R^2+\eta^2 t^2]/\alpha Q$$

$$P[x|y] = \frac{e^{-\frac{1}{2}[x-Ry-(\eta t/\alpha)\;\mathrm{sgn}(y)]^2/(Q-R^2)}}{\sqrt{2\pi(Q-R^2)}} \tag{5.20}$$

(this solution is unique). If we calculate the generalization error and the training error from (5.19) and (5.20), respectively, we recover the exact expressions

$$E_{\mathrm{g}} = \frac{1}{\pi}\arccos\left[\frac{R_0+\eta t\sqrt{\frac{2}{\pi}}}{\sqrt{Q_0+2\eta t R_0\sqrt{\frac{2}{\pi}}+\eta^2 t^2\left[\frac{1}{\alpha}+\frac{2}{\pi}\right]}}\right] \tag{5.21}$$

$$E_{\mathrm{t}} = \frac{1}{2} - \frac{1}{2}\int Dy\;\mathrm{erf}\left[\frac{|y|[R_0+\eta t\sqrt{\frac{2}{\pi}}]+\frac{\eta t}{\alpha}}{\sqrt{2[Q_0-R_0^2+\frac{\eta^2 t^2}{\alpha}]}}\right] \tag{5.22}$$

Comparison of (5.19,5.20) with (5.14,5.15) shows that for batch Hebbian learning our theory is fully exact. This is not a big feat as far as Q and R (and thus E_{g}) are concerned, whose determination did not require knowing the function $\Phi[x,y]$. The fact that our theory also gives the exact values for $P[x|y]$ and E_{t}, however, is less trivial, since here the disordered nature of the learning dynamics, leading to non-Gaussian distributions, is truly relevant.

5.5 On-Line Hebbian Learning

We next insert the Hebbian recipe $\mathcal{G}[x,y]=\text{sgn}[y]$ into the on-line equations (4.28,4.29). Direct analytical solution of these equations, or a demonstration that they are solved by the exact result (5.12,5.13), although not ruled out, has not yet been achieved. The reason is that here one has locally Gaussian field distributions only in special limits. Numerical solution is straightforward, but has not yet been carried out. For small learning rates the on-line equations reduce to the batch ones, so we know that in first order in η our on-line equations are exact (for any α, t). We now show that the predictions of our theory are fully exact (i) for Q, R and $E_{\rm g}$, (ii) for the first moment (5.16) of the conditional field distribution, and (iii) for all order parameters in the stationary state. At intermediate times we construct an approximate solution of our equations in order to obtain predictions for $P[x|y]$ and $E_{\rm t}$.

As before we choose a Gaussian initial field distribution. Many (but not all) of our previous simplifications still hold, e.g.

$$U=0, \qquad V=\langle x\ \text{sgn}(y)\rangle, \qquad W=\sqrt{2/\pi}, \qquad Z=1$$

(Z did not occur in the batch equations). Thus for on-line learning we find:

$$\frac{d}{dt}Q=2\eta V+\eta^2 \qquad\qquad \frac{d}{dt}R=\eta\sqrt{2/\pi}$$

The previous derivation of the identities $\frac{d}{dt}V=\eta/\alpha+2\eta/\pi$ and $V_0=R_0\sqrt{2/\pi}$ still applies (just replace the batch diffusion equation by the on-line one), but the resultant expression for Q is different. Here we obtain:

$$Q=Q_0+2\eta t R_0\sqrt{\frac{2}{\pi}}+\eta^2 t+\eta^2 t^2\left[\frac{1}{\alpha}+\frac{2}{\pi}\right] \qquad\qquad R=R_0+\eta t\sqrt{\frac{2}{\pi}} \tag{5.23}$$

Comparing (5.23) with (5.12) reveals that also for on-line Hebbian learning our theory is exact with regard to Q and R, and thus also with regard to $E_{\rm g}$. Upon using $V=\eta t/\alpha+R\sqrt{2/\pi}$, the on-line diffusion equation simplifies to

$$\frac{d}{dt}P[x|y]=\frac{1}{\alpha}\Big\{P[x-\eta\ \text{sgn}(y)|y]-P[x|y]\Big\}-\eta y\sqrt{\frac{2}{\pi}}\frac{\partial}{\partial x}P[x|y]+\frac{1}{2}\eta^2\frac{\partial^2}{\partial x^2}P[x|y]$$
$$-\frac{\eta^2 t}{\alpha}\frac{\partial}{\partial x}\Big\{P[x|y]\Phi[x,y]\Big\}$$

Multiplication of this equation by x followed by integration over x, together with the general properties $\int dx\ \{P[x|y]\Phi[x,y]\}=0$ and $\int dx\ xP_0[x|y]=R_0 y$, gives us the average of the conditional distribution $P[x|y]$ at any time:

$$\overline{x}(y)=\int dx\ xP[x|y]=Ry+\frac{\eta t}{\alpha}\ \text{sgn}[y]$$

Comparison with (5.16) shows also this prediction to be correct.

We now turn to observables which involve more detailed knowledge of the function $\Phi[x,y]$. Our result for $\overline{x}(y)$ and the identity $\langle x\rangle_\star = B^{-1}\frac{\partial}{\partial z}\log\hat{M}[iBz|y]$ allow us to rewrite all remaining equations in Fourier representation, i.e. in terms of $\hat{P}[k|y]=\int dx\ e^{-ikx}P[x|y]$ and $\hat{M}[k|y]=\int dx\ e^{-ikx}M[x|y]$:

$$\frac{d}{dt}\log\hat{P}[k|y] = \frac{1}{\alpha}\left[e^{-i\eta k\ \mathrm{sgn}(y)}-1\right] - i\eta k y\sqrt{\frac{2}{\pi}} - \frac{1}{2}\eta^2k^2$$
$$-\frac{ik\eta^2 t}{\alpha}\left[\hat{P}[k|y]\sqrt{qQ-R^2}\right]^{-1}\int Dz\ z\ \frac{\hat{M}[k+iBz|y]}{\hat{M}[iBz|y]} \qquad (5.24)$$

with $\log\hat{P}_0[k|y] = -ikR_0y-\frac{1}{2}k^2(Q_0-R_0^2)$, and with the saddle-point equations

$$\hat{P}[k|y] = \int Dz\ \frac{\hat{M}[k+iBz|y]}{\hat{M}[iBz|y]} \qquad (5.25)$$

$$\frac{\eta^2t^2}{\alpha^2} + \int Dy\int dx\ P[x|y][x-\overline{x}(y)]^2 + (1-\frac{1}{\alpha})(qQ-R^2)$$
$$= \left[2Q(1-q)+\frac{1}{B^2}\right]\int DyDz\ \frac{\partial^2}{\partial z^2}\log\hat{M}[iBz|y] \qquad (5.26)$$

Due to the fields x growing linearly with time (see our expression for $\overline{x}(y)$) the equations (5.24,5.26,5.25) cannot have proper $t\to\infty$ limits. To extract asymptotic properties we have to turn to the rescaled distribution $\hat{Q}[k|y]=\hat{P}[k/t|y]$. We define $v(y)=(\eta/\alpha)\,\mathrm{sgn}(y)+\eta y\sqrt{2/\pi}$. Careful integration of (5.24), followed by inserting $k\to k/t$ and by taking the limit $t\to\infty$, produces:

$$\log\hat{Q}_\infty[k|y] = -ikv(y) - \frac{i\eta^2k}{\alpha}\int_0^1 du\ \lim_{t\to\infty}\frac{t}{\sqrt{qQ-R^2}}\int Dz\ z\ \frac{\hat{M}[uk/t+iBz|y]}{\hat{Q}_\infty[uk|y]\hat{M}[iBz|y]} \qquad (5.27)$$

with the functional saddle-point equation

$$\hat{Q}[k|y] = \int Dz\ \frac{\hat{M}[k/t+iBz|y]}{\hat{M}[iBz|y]} \qquad (5.28)$$

The rescaled asymptotic system (5.27,5.28) admits the solution

$$\hat{Q}[k|y] = e^{-ikv(y)-\frac{1}{2}k^2\tilde{\Delta}^2}, \qquad \hat{M}[k|y] = e^{-ik\overline{x}(y)-\frac{1}{2}k^2\tilde{\sigma}^2t}$$

with the asymptotic values of B, $\tilde{\Delta}$, $\tilde{\sigma}$ and q determined by solving

$$\tilde{\Delta} = B\tilde{\sigma}^2 \qquad \tilde{\Delta} = \frac{\eta^2}{\alpha}\lim_{t\to\infty}\frac{t}{\sqrt{qQ-R^2}} \qquad B = \lim_{t\to\infty}\frac{\sqrt{qQ-R^2}}{Q(1-q)}$$

$$\eta^2/\alpha^2 + \tilde{\Delta}^2 + (1-\alpha^{-1})\lim_{t\to\infty}(qQ-R^2)/t^2 = 2B^2\tilde{\sigma}^2\lim_{t\to\infty}Q(1-q)/t$$

Inspection shows that these four asymptotic equations are solved by

$$\lim_{t\to\infty}\tilde{\Delta}=\eta/\sqrt{\alpha},\qquad \lim_{t\to\infty}q=1$$

so that

$$\lim_{t\to\infty}\hat{P}_t[k/t]=e^{-ik\eta\left[\alpha^{-1}\ \mathrm{sgn}(y)+y\sqrt{2/\pi}\right]-\frac{1}{2}\eta^2k^2/\alpha} \tag{5.29}$$

Comparison with (5.12,5.13) shows that this prediction (5.29) is again exact. Thus the same is true for the asymptotic training error.

Finally, in order to arrive at predictions with respect to $P[x|y]$ and $E_{\rm t}$ for intermediate times (without rigorous analytical solution of the functional saddle-point equation), and in view of the locally Gaussian form of the field distribution both at $t=0$ and at $t=\infty$, we can approximate $P[x|y]$ and $M[x|y]$ by simple locally Gaussian distributions at any time:

$$P[x|y]=\frac{e^{-\frac{1}{2}[x-\overline{x}(y)]^2/\Delta^2}}{\Delta\sqrt{2\pi}},\qquad M[x|y]=\frac{e^{-\frac{1}{2}[x-\overline{x}(y)]^2/\sigma^2}}{\sigma\sqrt{2\pi}} \tag{5.30}$$

with the (exact) first moments $\overline{x}(y)=Ry+\eta t\alpha^{-1}\,\mathrm{sgn}(y)$, and with the variance Δ^2 self-consistently given by the solution of:

$$\Delta^2=\sigma^2+B^2\sigma^4\qquad B=\frac{\sqrt{qQ-R^2}}{Q(1-q)}\qquad \frac{d}{dt}\Delta^2=\frac{\eta^2}{\alpha}+\eta^2+\frac{2\eta^2t\sigma^2}{\alpha Q(1-q)}$$

$$\alpha\Delta^2+\frac{\eta^2t^2}{\alpha}+(qQ-R^2)(\alpha-1)=\alpha\sigma^2\left[2\frac{qQ-R^2}{Q(1-q)}+1\right]$$

The solution of the above coupled equations behaves as

$$\Delta^2=Q-R^2+\eta^2t/\alpha+\mathcal{O}(t^3)\qquad (t\to 0)$$
$$\Delta^2/(Q-R^2)=\mathcal{O}(t^{-1})\qquad (t\to\infty)$$

for short and long times, respectively (note $Q-R^2\sim t^2$ as $t\to\infty$). Thus we obtain a simple approximate solution of our equations, which extrapolates between exact results at the temporal boundaries $t=0$ and $t=\infty$, by putting

$$\Delta^2=Q-R^2+\eta^2t/\alpha$$

with Q and R given by our previous exact result (5.23), which results in

$$E_{\rm g}=\frac{1}{\pi}\arccos\left[\frac{R}{\sqrt{Q}}\right]\qquad E_{\rm t}=\frac{1}{2}-\frac{1}{2}\int Dy\ \mathrm{erf}\left[\frac{|y|R+\eta t/\alpha}{\Delta\sqrt{2}}\right] \tag{5.31}$$

We can also calculate the student field distribution $P(x)=\int Dy\,P[x|y]$, giving

$$P(x)=\frac{e^{-\frac{1}{2}[x+\frac{\eta t}{\alpha}]^2/(\Delta^2+R^2)}}{2\sqrt{2\pi(\Delta^2+R^2)}}\left[1-\mathrm{erf}\left(\frac{R[x+\eta t/\alpha]}{\Delta\sqrt{2(\Delta^2+R^2)}}\right)\right]$$
$$+\frac{e^{-\frac{1}{2}[x-\frac{\eta t}{\alpha}]^2/(\Delta^2+R^2)}}{2\sqrt{2\pi(\Delta^2+R^2)}}\left[1+\mathrm{erf}\left(\frac{R[x-\eta t/\alpha]}{\Delta\sqrt{2(\Delta^2+R^2)}}\right)\right] \tag{5.32}$$

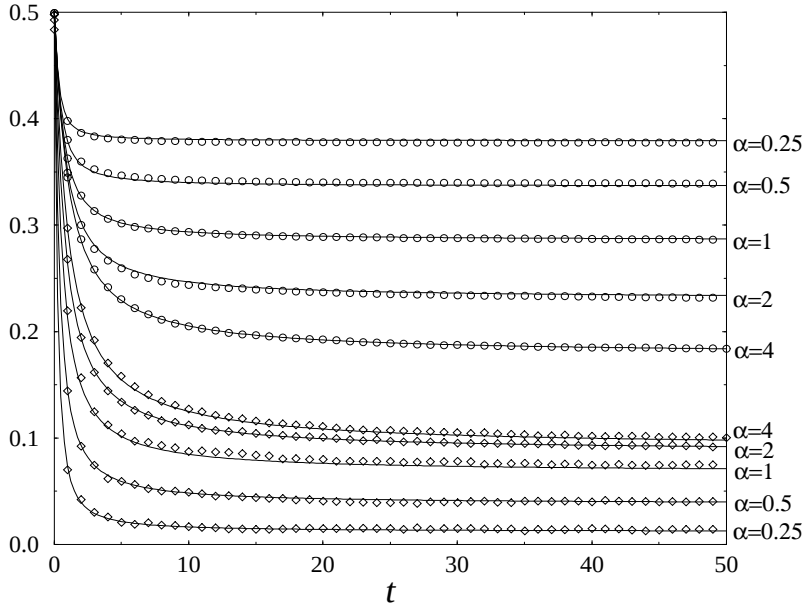

Fig. 4: On-line Hebbian learning, simulations versus theoretical predictions, for $\alpha \in \{0.25, 0.5, 1.0, 2.0, 4.0\}$ ($N = 10,000$). Upper curves: generalization errors as functions of time. Lower curves: training errors as functions of time. Circles: simulation results for $E_{\rm g}$; diamonds: simulation results for $E_{\rm t}$. Solid lines: corresponding predictions of dynamical replica theory.

5.6 Comparison with Simulations

In Fig. 4 we compare the predictions for the generalization and training errors (5.31) of the approximate solution of our equations with the results obtained from numerical simulations of on-line Hebbian learning for $N = 10,000$ (initial state: $Q_0 = 1$, $R_0 = 0$; learning rate: $\eta = 1$). All curves show excellent agreement between theory and experiment. For $E_{\rm g}$ this is guaranteed by the exactness of our theory for Q and R; the agreement found for $E_{\rm t}$ is more surprising, in that these predictions are obtained from a simple approximation of the solution of our equations. We also compare the theoretical predictions made for the distribution $P[x|y]$ with the results of numerical simulations. This is done in Fig. 5, where we show the fields as observed at time $t = 50$ in simulations ($N = 10,000$, $\eta = 1$, $R_0 = 0$, $Q_0 = 1$) of on-line Hebbian learning, for three different values of α. In the same figure we draw (as dashed lines) the theoretical prediction (5.16) for the y-dependent average of the conditional x-distribution $P[x|y]$. Finally we compare the student field distribution $P(x)$, as observed in simulations of on-line Hebbian learning ($N = 10,000$, $\eta = 1$, $R_0 = 0$, $Q_0 = 1$) with our prediction (5.32). The result is shown in Fig. 6, for $\alpha \in \{4, 1, 0.25\}$. In all cases the agreement between theory and experiment, even for the approximate solution of our equations, is quite satisfactory.

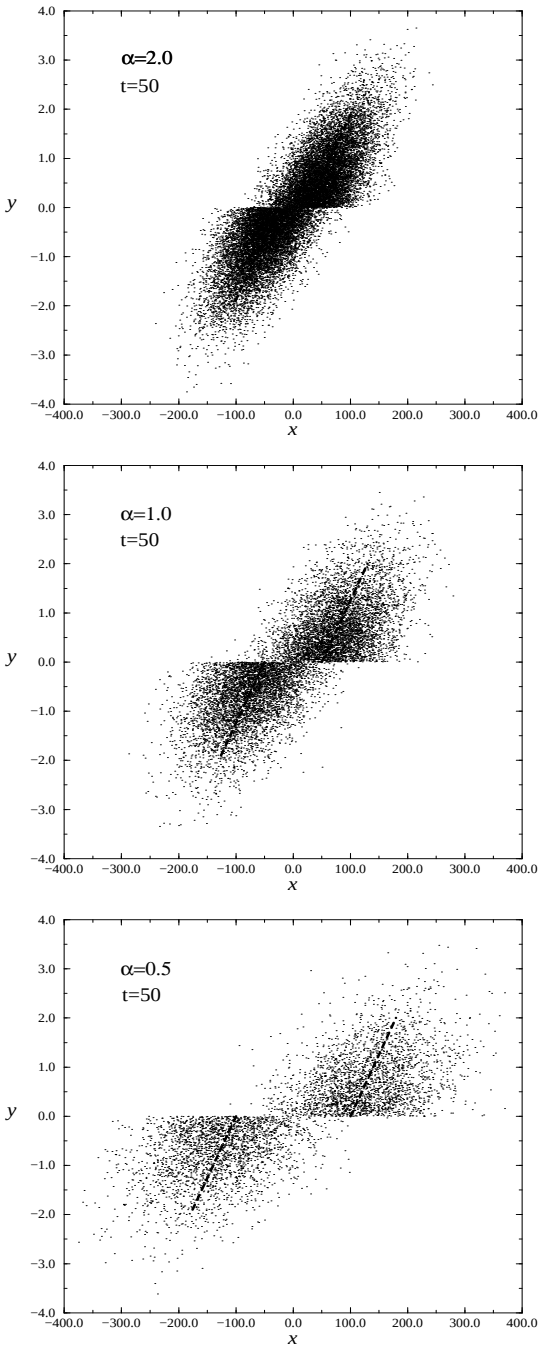

Fig. 5: Comparison between simulation results for on-line Hebbian learning (system size $N = 10,000$) and dynamical replica theory, for $\alpha \in \{0.5, 1.0, 2.0\}$. Dots: local fields $(x, y) = (\boldsymbol{J}\cdot\boldsymbol{\xi}, \boldsymbol{B}\cdot\boldsymbol{\xi})$ (calculated for questions in the training set), at time $t = 50$. Dashed lines: conditional average of student field x as a function of y, as predicted by the theory, $\overline{x}(y) = Ry + (\eta t/\alpha)\ \mathrm{sgn}(y)$.

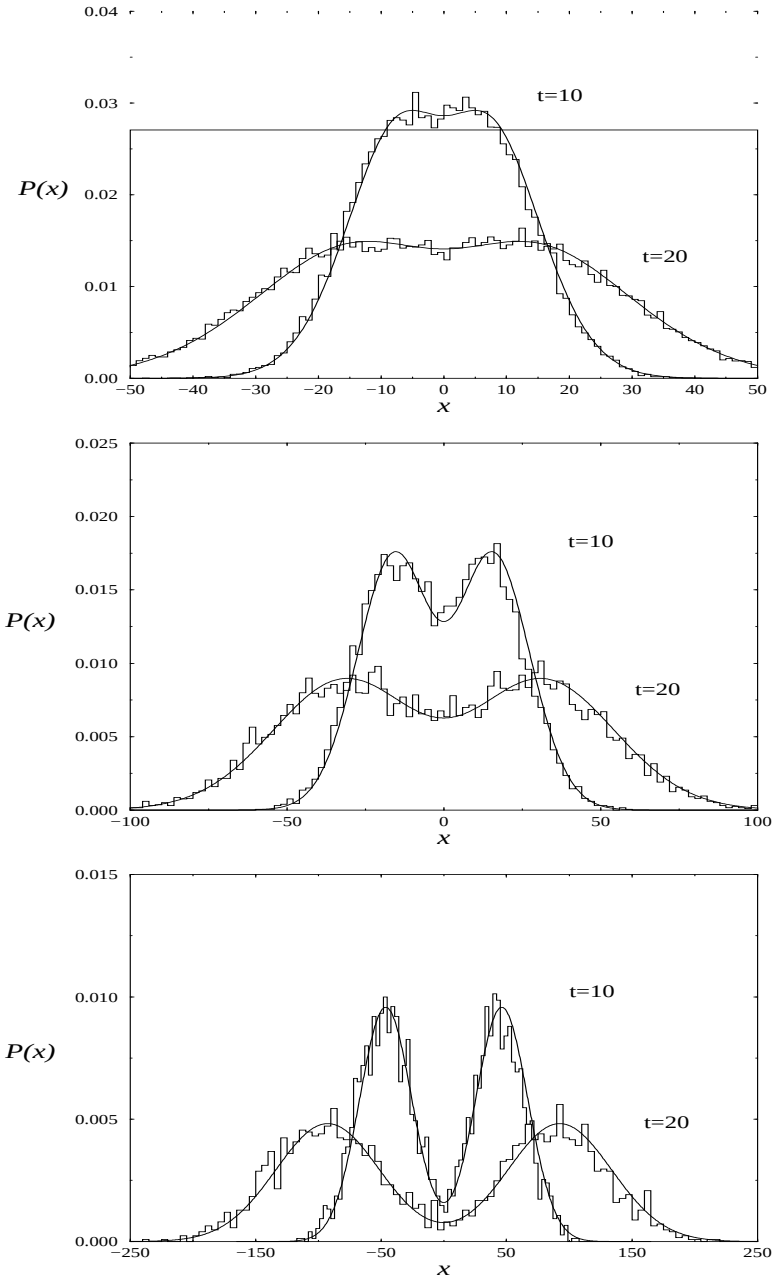

Fig. 6: Simulations of Hebbian on-line learning with $N = 10,000$. Histograms: student field distributions measured at $t = 10$ and $t = 20$. Lines: theoretical predictions for student field distributions. $\alpha = 4$ (upper), $\alpha = 1$ (middle), $\alpha = 0.25$ (lower).

6 Discussion

In this paper we have shown how the formalism of dynamical replica theory (e.g. Coolen et al, 1996) can be used successfully to build a general theory with which to predict the evolution of the relevant macroscopic performance measures for supervised (on-line and batch) learning in layered neural networks with randomly composed but restricted training sets (i.e. for finite $\alpha = p/N$), where the student fields are no longer described by Gaussian distributions, and where the more traditional and familiar statistical mechanical formalism consequently breaks down. For simplicity and transparency we have restricted ourselves to single-layer systems and realizable tasks. In our approach the joint field distribution $P[x, y]$ for student and teacher fields is itself taken to be a dynamical order parameter, in addition to the more conventional observables Q and R; from this order parameter set $\{Q, R, P\}$, in turn, immediately follow the generalization error $E_{\rm g}$ and the training error $E_{\rm t}$. This then results, following the prescriptions of dynamical replica theory[4], in a diffusion equation for $P[x, y]$, which we have evaluated by making the replica-symmetric ansatz in the saddle-point equations. This diffusion equation is found to have Gaussian solutions only for $\alpha \to \infty$; in the latter case we indeed recover correctly from our theory the more familiar formalism of infinite training sets, with (in the $N \to \infty$ limit) closed equations for Q and R only. For finite α our theory is by construction exact if for $N \to \infty$ the dynamical order parameters $\{Q, R, P\}$ obey closed, deterministic equations, which are self-averaging (i.e. independent of the microscopic realization of the training set). If this is not the case, our theory is an approximation.

We have worked out our equations explicitly for the special case of Hebbian learning, where the availability of exact results, derived directly from the microscopic equations, allows us to perform a critical test of our theory [5]. For batch Hebbian learning we can demonstrate explicitly that our theory is fully exact. For on-line Hebbian learning, on the other hand, proving or disproving full exactness requires solving a non-trivial functional saddle-point equation analytically, which we have not yet been able to do. Nevertheless we can prove that our theory is exact (i) with respect to its predictions for Q, R and $E_{\rm g}$, (ii) with respect to the first moment of the conditional field distribution $P[x|y]$, and (iii) in the stationary state. In order to also generate predictions for intermediate times we have constructed an approximate solution of our equations, which is found to describe the results of performing numerical simulations of on-line Hebbian learning quite satisfactorily.

[4] The reason why replicas are inevitable (unless we are willing to pay the price of having observables with two time arguments, and turn to path integrals) is the necessity, for finite α, to average the macroscopic equations over all possible realizations of the training set.

[5] Such exact results can only be obtained for Hebbian-type rules, where the dependence of the updates $\Delta \boldsymbol{J}(t)$ on the weights $\boldsymbol{J}(t)$ is trivial or even absent (a decay term at most), whereas our present theory generates macroscopic equations for arbitrary learning rules.

The present study represents only a first step; many extensions, applications and generalizations can be carried out (most of which are already under way). Firstly, our theory would greatly simplify if we could find an explicit solution of the functional saddle-point equation, enabling us to express the function $\Phi[x, y]$ directly in terms of our order parameters. The benefits of such a solution will become even greater when we apply our theory to more sophisticated learning rules, such as to perceptron or AdaTron learning, or to learning in multi-layer networks (which run the risk of requiring a serious amount of CPU time). Yet another direction is the inclusion of unlearnable tasks, such as those generated by noisy teachers. At a more fundamental level one could explore the potential of (dynamic) replica symmetry breaking (by calculating the AT-surface, signaling instability of the replica symmetric solution with respect to replicon fluctuations), or one could improve the built-in accuracy of our theory by adding new observables to the present set (such as the Green's function $\mathcal{A}[x, y; x', y']$ itself). Finally it would be interesting to see the connection between the present formalism and a suitable adaptation of the work by Horner (1992), based on generating functionals and path integrals, to the processes studied in this paper (with non-binary weights).

Acknowledgement:
DS acknowledges support by EPSRC Grant GR/L52093.

References

Barber D., Saad D. and Sollich P. (1996), *Europhys. Lett.* **34**, 151

Biehl M. and Schwarze H. (1992), *Europhys. Lett.* **20**, 733

Biehl M. and Schwarze H. (1995), *J. Phys. A: Math. Gen.* **28**, 643

Coolen A.C.C., Laughton S.N. and Sherrington D. (1996), *Phys. Rev. B* **53**, 8184

Coolen, A.C.C and Saad D. (1998), in preparation.

Horner H. (1992a), *Z. Phys. B* **86**, 291

Horner H. (1992b), *Z. Phys. B* **87**, 371

Kinouchi O. and Caticha N. (1992), *J. Phys. A: Math. Gen.* **25**, 6243

Kinzel W. and Rujan P. (1990), *Europhys. Lett.* **13**, 473

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

Mace C.W.H. and Coolen A.C.C (1998b), in preparation

Mézard M., Parisi G. and Virasoro M.A. (1987), *Spin-Glass Theory and Beyond* (Singapore: World Scientific)

Rae H.C., Sollich P. and Coolen A.C.C. (1998), in preparation

Saad D. and Coolen, A.C.C (1998), in preparation

Saad D. and Solla S. (1995), *Phys. Rev. Lett.* **74**, 4337

Sollich P. and Barber D. (1997), to be published in Proc. NIPS*97