Dynamics of Learning with Restricted Training Sets

A.C.C. Coolen
Department of Mathematics
King's College London
Strand, London WC2R 2LS, UK

D. Saad
The Neural Computing Research Group
Aston University
Birmingham B4 7ET, UK

April 27th 2000

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 probability distributions and the learning dynamics is of a spin-glass nature, with the composition of the training set playing the role of quenched disorder. We show how dynamical replica theory can be used to predict the evolution of macroscopic observables, including the two relevant performance measures (training error and generalization error), incorporating the old formalism developed for complete training sets in the limit $\alpha = p/N \to \infty$ as a special case. For simplicity we restrict ourselves in this paper to single-layer networks and realizable tasks. In the case of (on-line and batch) Hebbian learning, where a direct exact solution is possible, we show that our theory provides exact results at any time in many different verifiable cases. For non-Hebbian learning rules, such as Perceptron and AdaTron, we find very good agreement between the predictions of our theory and numerical simulations. Finally, we derive three approximation schemes aimed at eliminating the need to solve a functional saddle-point equation at each time step, and assess their performance. The simplest of these schemes leads to a fully explicit and relatively simple non-linear diffusion equation for the joint field distribution, which already describes the learning dynamics surprisingly well over a wide range of parameters.

PACS: 87.10.+e, 02.50.-r, 05.20.-y

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 of the learning process a set of closed laws describing the evolution of suitably chosen macroscopic observables (dynamic order parameters), in the limit of an infinite system size (e.g. [1, 2, 3, 4, 5]. A recent review and more extensive guide to the relevant references can be found in [6]. A preliminary presentation of some of the present results was given in [7]. The main successful procedure developed so far is built on the following four cornerstones:

- *The task to be learned by the network 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 weight updates made.* Each example presented to the system 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 [8], 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.

Here 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. $p = \alpha N$ with $0 < \alpha < \infty$. In this regime 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 individual 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 indeed 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 complete 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. [9], 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 in [10] for perceptrons with binary weights. The only cases where explicit and relatively simple solutions can be obtained, even for restricted training sets, are those where linear learning rules are used, such as [11] or [12].

In this paper we show how the formalism of dynamical replica theory (see e.g. [13]) 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 $P[x,y]$ 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 in the present paper to single-layer systems and noise-free teachers. Application and generalization of our methods to multi-layer systems [14] and learning scenarios involving 'noisy' teachers [15] are presently under way.

Our paper is organized as follows. We first derive a Fokker-Planck equation describing the evolution of arbitrary mean-field observables for $N \to \infty$. This allows us to identify the conditions for the latter to be described by closed deterministic laws. We then choose as our observables the joint field distribution $P[x,y]$, in addition to (the traditional ones) Q and R, and show that this set $\{Q,R,P\}$ 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 an Appendix, so that they can be skipped by those primarily interested in results. We summarize the final replica-symmetric macroscopic theory and its notational conventions, discuss some of its general properties, and show how in the limit $\alpha \to \infty$ (infinite training sets) the equations of the conventional theory are recovered. We then apply our general theory to various different specific choices of learning rules. One of these, (on-line and batch) Hebbian learning, provides an excellent benchmark test for our theory, since for this simple rule exact solutions are known, even for the regime of restricted training sets [12]. We find that our theory is fully exact for batch execution, and that it succeeds in predicting exactly the evolution of several macroscopic observables, including the generalisation error and moments of the joint field distribution for student and teacher fields, in the on-line case (although here full exactness is difficult to assess, and not a priori guaranteed). For non-Hebbian error-correcting learning rules, such as on-line and batch versions of Perceptron learning and AdaTron learning, no exact solutions are known at present with which to confront our theory; instead we here compare the predictions (with regard to the evolution of training- and generalization errors and the joint field distribution) of the full theory, as well as of a number of simple approximations of our equations, with the results of carrying out extensive numerical simulations in large (size $N = 10,000$) neural networks. We find, surprisingly, that even the simplest of these approximations, which does not require solving any saddle point equations and takes the form of a fully explicit non-linear diffusion equation for the joint field distributions $P[x,y]$, describes the simulation experiments remarkably well. Employing the more sophisticated (and thereby more CPU intensive) approximations, or, at the other end of the spectrum, a numerical solution of the full macroscopic theory, leads to increasingly accurate quantitative predictions for the evolution of the relevant macroscopic observables of the learning process, with deviations between theory and numerical experiment which are of the order of magnitude of the finite size effects in the simulations. We close our paper with a discussion of the strengths and weaknesses of the approach used, and an outlook on future work on the dynamics of learning with restricted training sets, involving the present and possibly other formalisms.

2 From Microscopic to Macroscopic Laws

2.1 Definitions

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

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

It tries to emulate the operation of a teacher perceptron, which is assumed to operate a similar rule, characterized by a given (fixed) weight vector $\boldsymbol{B} \in \Re^N$:

$$T : \{-1,1\}^N \to \{-1,1\} \qquad\qquad T(\boldsymbol{\xi}) = \ \mathrm{sgn}\,[\boldsymbol{B} \cdot \boldsymbol{\xi}] \tag{2}$$

In order to improve its performance, the student perceptron modifies its weight vector $\boldsymbol{J}$ according 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 $T(\boldsymbol{\xi})$.

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

$$\tilde{D} = \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p\} \qquad\qquad p = \alpha N \qquad\qquad \boldsymbol{\xi}^\mu \in D \ \text{ for all } \mu \tag{3}$$

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:

$$\begin{array}{ll} \text{on-line}: & \boldsymbol{J}(m+1) = \boldsymbol{J}(m) + \frac{\eta}{N} \ \boldsymbol{\xi}(m) \ \mathcal{G}\left[\boldsymbol{J}(m) \cdot \boldsymbol{\xi}(m), \boldsymbol{B} \cdot \boldsymbol{\xi}(m)\right] \\ \text{batch}: & \boldsymbol{J}(m+1) = \boldsymbol{J}(m) + \frac{\eta}{N} \ \langle \boldsymbol{\xi} \ \mathcal{G}\left[\boldsymbol{J}(m) \cdot \boldsymbol{\xi}, \boldsymbol{B} \cdot \boldsymbol{\xi}\right] \rangle_{\tilde{D}} \end{array} \tag{4}$$

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:

$$E_{\rm t}(\boldsymbol{J}) = \langle \theta[-(\boldsymbol{J} \cdot \boldsymbol{\xi})(\boldsymbol{B} \cdot \boldsymbol{\xi})] \rangle_{\tilde{D}} \qquad\qquad E_{\rm g}(\boldsymbol{J}) = \langle \theta[-(\boldsymbol{J} \cdot \boldsymbol{\xi})(\boldsymbol{B} \cdot \boldsymbol{\xi})] \rangle_{D} \ . \tag{5}$$

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.

We next convert the dynamical laws (4) 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 (4) can be written in the general Markovian form

$$\hat{p}_{m+1}(\boldsymbol{J}) = \int d\boldsymbol{J}' \ W[\boldsymbol{J};\boldsymbol{J}'] \ \hat{p}_m(\boldsymbol{J}') \ , \tag{6}$$

with the transition probabilities

$$\begin{array}{ll} \text{on-line}: & W[\boldsymbol{J};\boldsymbol{J}'] = \langle \delta\left[\boldsymbol{J} - \boldsymbol{J}' - \frac{\eta}{N} \ \boldsymbol{\xi} \ \mathcal{G}\left[\boldsymbol{J}' \cdot \boldsymbol{\xi}, \boldsymbol{B} \cdot \boldsymbol{\xi}\right]\right] \rangle_{\tilde{D}} \\ \text{batch}: & W[\boldsymbol{J};\boldsymbol{J}'] = \delta\left[\boldsymbol{J} - \boldsymbol{J}' - \frac{\eta}{N} \langle \boldsymbol{\xi} \ \mathcal{G}\left[\boldsymbol{J}' \cdot \boldsymbol{\xi}, \boldsymbol{B} \cdot \boldsymbol{\xi}\right] \rangle_{\tilde{D}}\right] \end{array} \tag{7}$$

We 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

$$\pi_m(t) = \frac{1}{m!}(Nt)^m e^{-Nt} \ . \tag{8}$$

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:

$$p_t(\boldsymbol{J}) = \sum_{m \geq 0} \pi_m(t) \ \hat{p}_m(\boldsymbol{J}) \ . \tag{9}$$

This distribution obeys a simple differential equation, which immediately follows from the pleasant properties of (8) under temporal differentiation:

$$\frac{d}{dt} \ p_t(\boldsymbol{J}) = N \int d\boldsymbol{J}' \ \left\{ W[\boldsymbol{J};\boldsymbol{J}'] - \delta[\boldsymbol{J} - \boldsymbol{J}'] \right\} \ p_t(\boldsymbol{J}') \ . \tag{10}$$

So far no approximations have been made, equation (10) is exact for any N. It is the equivalent of the master equation often introduced to define the dynamics of spin systems.

2.2 Derivation of Macroscopic Fokker-Planck Equation

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

$$P_t(\boldsymbol{\Omega}) = \int d\boldsymbol{J}\ p_t(\boldsymbol{J})\delta\left[\boldsymbol{\Omega} - \boldsymbol{\Omega}[\boldsymbol{J}]\right] \tag{11}$$

Its time derivative immediately follows from that in (10):

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

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

This then can be written in the standard form

$$\frac{d}{dt}P_t(\boldsymbol{\Omega}) = \int d\boldsymbol{\Omega}'\ \mathcal{W}_t[\boldsymbol{\Omega};\boldsymbol{\Omega}']P_t(\boldsymbol{\Omega}') \tag{12}$$

where

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

If we now insert the relevant expressions (7) for $W[\boldsymbol{J};\boldsymbol{J}']$ we can perform the $\boldsymbol{J}$-integrations, and obtain results given in terms of so-called sub-shell averages, which are defined as

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

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

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

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

We now insert integral representations for the δ-distributions. The observables $\boldsymbol{\Omega}[\boldsymbol{J}] \in \Re^k$ are assumed to be $\mathcal{O}(1)$ each, and finite in number (i.e. $k \ll N$):

$$\delta[\boldsymbol{\Omega}-\boldsymbol{Q}] = \int\frac{d\hat{\boldsymbol{\Omega}}}{(2\pi)^k}e^{i\hat{\boldsymbol{\Omega}}\cdot[\boldsymbol{\Omega}-\boldsymbol{Q}]} \tag{13}$$

which gives for our two learning scenario's:

$$\mathcal{W}_t^{\rm onl}[\boldsymbol{\Omega};\boldsymbol{\Omega}'] = \int\frac{d\hat{\boldsymbol{\Omega}}}{(2\pi)^k}e^{i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}}\ N\left\langle\langle e^{-i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}+\frac{\eta}{N}\boldsymbol{\xi}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]]}\rangle_{\bar{D}} - e^{-i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}]}\right\rangle_{\boldsymbol{\Omega}';t} \tag{14}$$

$$\mathcal{W}_t^{\rm bat}[\boldsymbol{\Omega};\boldsymbol{\Omega}'] = \int\frac{d\hat{\boldsymbol{\Omega}}}{(2\pi)^k}e^{i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}}\ N\left\langle e^{-i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}+\frac{\eta}{N}\langle\boldsymbol{\xi}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}]} - e^{-i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}]}\right\rangle_{\boldsymbol{\Omega}';t} \tag{15}$$

Still no approximations have been made. The above two expressions differ only in at which stage the averaging over the training set occurs.

In expanding equations (14,15) 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. We therefore inspect more closely the usual Taylor expansions:

$$F[\boldsymbol{J}+\boldsymbol{k}]-F[\boldsymbol{J}]=\sum_{\ell\geq 1}\frac{1}{\ell!}\sum_{i_1=1}^{N}\cdots\sum_{i_\ell=1}^{N}k_{i_1}\cdots k_{i_\ell}\frac{\partial^\ell F[\boldsymbol{J}]}{\partial J_{i_1}\cdots\partial J_{i_\ell}}\ .$$

If we assess how derivatives with respect to individual components J_i scale for mean-field 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:

$$F[\boldsymbol{J}]=\mathcal{O}(N^0),\qquad\qquad \frac{\partial^\ell F[\boldsymbol{J}]}{\partial J_{i_1}\cdots\partial J_{i_\ell}}=\mathcal{O}\left(|\boldsymbol{J}|^{-\ell}N^{\frac{1}{2}\ell-d}\right)\qquad (N\to\infty) \tag{16}$$

in which d is the number of different elements in the set $\{i_1,\ldots,i_\ell\}$. For simple mean-field observables we can now estimate the scaling of the various terms in the Taylor expansion. However, we will find that for restricted training sets not all relevant observables will have the properties (16). In particular, the joint distribution of student and teacher fields will, for on-line learning, have a contribution for which all terms in the Taylor series will have to be summed, giving rise to an additional term $\Delta[\boldsymbol{J};\boldsymbol{k}]$ [1]. The latter type of more *general* mean-field observables will have to be defined via the identities

$$F[\boldsymbol{J}+\boldsymbol{k}]-F[\boldsymbol{J}]=\Delta[\boldsymbol{J};\boldsymbol{k}]+\sum_i k_i\frac{\partial F[\boldsymbol{J}]}{\partial J_i}+\frac{1}{2}\sum_{ij}k_ik_j\frac{\partial^2F[\boldsymbol{J}]}{\partial J_i\partial J_j}+\sum_{\ell\geq 3}\mathcal{O}\left(\left[\frac{|\boldsymbol{k}|}{|\boldsymbol{J}|}\right]^\ell\right) \tag{17}$$

$$F[\boldsymbol{J}]=\mathcal{O}(N^0),\qquad\qquad \Delta[\boldsymbol{J};\boldsymbol{k}]=\mathcal{O}\left(|\boldsymbol{k}|^2/|\boldsymbol{J}|^2\right) \tag{18}$$

(in the assessment of the order of the remainder terms of (17) 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 expand our macroscopic equations (14,15) for large N and finite times, restricting ourselves from now on to mean-field observables in the sense of (17,18). One of our observables we choose to be $\boldsymbol{J}^2$. In the present problem the shifts $\boldsymbol{k}$, being either $\frac{\eta}{N}\boldsymbol{\xi}\,G[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]$ or $\frac{\eta}{N}\langle\boldsymbol{\xi}\,\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}$, scale as $|\boldsymbol{k}|=\mathcal{O}(N^{-\frac{1}{2}})$. Consequently:

$$e^{-i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}+\boldsymbol{k}]}=e^{-i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}]}\left\{1-i\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Delta}\,[\boldsymbol{J};\boldsymbol{k}]-i\sum_i k_i\frac{\partial}{\partial J_i}(\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}])-\frac{i}{2}\sum_{ij}k_ik_j\frac{\partial^2}{\partial J_i\partial J_j}(\hat{\boldsymbol{\Omega}}\cdot\boldsymbol{\Omega}[\boldsymbol{J}])\right.$$

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

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 (14,15), that both types of dynamics are described by macroscopic laws with transition probability densities of the general form

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

which, due to (12) and for $N\to\infty$ and finite times, leads to a Fokker-Planck equation:

$$\frac{d}{dt}P_t(\mathbf{\Omega}) = -\sum_{\mu=1}^k \frac{\partial}{\partial\Omega_\mu}\left\{F_\mu[\mathbf{\Omega};t]P_t(\mathbf{\Omega})\right\} + \frac{1}{2}\sum_{\mu\nu=1}^k \frac{\partial^2}{\partial\Omega_\mu\partial\Omega_\nu}\left\{G_{\mu\nu}[\mathbf{\Omega};t]P_t(\mathbf{\Omega})\right\}\;. \tag{19}$$

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

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

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

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

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

Equation (19) 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

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

In the case of having 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.

2.3 Choice and Properties of Canonical Observables

We next apply the general results obtained so far to a specific set of observables, $\mathbf{\Omega}\to\{Q,R,P\}$, which are tailored to the problem at hand (note that we restrict ourselves to $\boldsymbol{J}^2=\mathcal{O}(1)$ and $\boldsymbol{B}^2=1$):

$$Q[\boldsymbol{J}] = \boldsymbol{J}^2, \qquad R[\boldsymbol{J}] = \boldsymbol{J}\cdot\boldsymbol{B}, \qquad P[x,y;\boldsymbol{J}] = \langle\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\;\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde D} \tag{20}$$

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 (19).

The observables (20) are indeed of the general mean-field type in the sense of (17,18). Insertion into the stronger condition (16) immediately shows this to be true for the scalar observables $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$ (they are simple mean field observables, for which the term (18) is absent). Verification of (17,18) 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 $p=\alpha N$ (where each realization of $\tilde{D}$ has equal probability) as $\langle\ldots\rangle_{\rm sets}$, this allows us to conclude

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

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 found to be a simple mean-field observable in the sense of (16).

The scaling properties of expansions or derivations of $P[x,y;\boldsymbol{J}]$ for a given training set $\tilde{D}$, however, need not be identical to those of its average over all training sets $\langle P[x,y;\boldsymbol{J}]\rangle_{\rm sets}$. Here we have to use the fact that $\tilde{D}$ has been composed in a random manner, as well as the specific form of the shifts $\boldsymbol{k}$ in $P[x,y;\boldsymbol{J}+\boldsymbol{k}]$ that occur for the two types of dynamics under consideration:

$$P[x,y;\boldsymbol{J}+\boldsymbol{k}]-P[x,y;\boldsymbol{J}]=\int\frac{d\hat{x}\;d\hat{y}}{(2\pi)^2}e^{i[x\hat{x}+y\hat{y}]}\frac{1}{p}\sum_{\mu=1}^{p}e^{-i\hat{x}\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu-i\hat{y}\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu}\left[e^{-i\hat{x}\boldsymbol{k}\cdot\boldsymbol{\xi}^\mu}-1\right]$$

All complications are caused by the dependence of $\boldsymbol{k}$ on the composition of the training set $\tilde{D}$, and would therefore have been absent in the $\alpha\to\infty$ case. This dependence will turn out to be harmless in the case of batch learning, where $\boldsymbol{k}=\frac{\eta}{N}\langle\boldsymbol{\xi}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}}$ is an average over $\tilde{D}$, but will have a considerable impact in the case of on-line learning, where $\boldsymbol{k}=\frac{\eta}{N}\boldsymbol{\xi}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]$ is proportional to an individual member of $\tilde{D}$. Working out the relevant expression for on-line learning gives

$$P[x,y;\boldsymbol{J}+\boldsymbol{k}^{\rm onl}]-P[x,y;\boldsymbol{J}]=\int\frac{d\hat{x}\;d\hat{y}}{(2\pi)^2}e^{i[x\hat{x}+y\hat{y}]}\frac{1}{p}\sum_{\mu=1}^{p}e^{-i\hat{x}\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu-i\hat{y}\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu}\left\{\delta_{\boldsymbol{\xi}^\mu\boldsymbol{\xi}}\left[e^{-i\eta\hat{x}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]}-1\right]\right.$$

$$\left.-[1-\delta_{\boldsymbol{\xi}^\mu,\boldsymbol{\xi}}]\left[\frac{i\eta\hat{x}}{N}(\boldsymbol{\xi}\cdot\boldsymbol{\xi}^\mu)\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]+\frac{\eta^2\hat{x}^2}{2N^2}(\boldsymbol{\xi}\cdot\boldsymbol{\xi}^\mu)^2\mathcal{G}^2[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]+\mathcal{O}(N^{-\frac{3}{2}})\right]\right\}$$

$$=\frac{1}{p}\int\frac{d\hat{x}d\hat{y}}{(2\pi)^2}e^{i[x\hat{x}+y\hat{y}]}e^{-i\hat{x}\boldsymbol{J}\cdot\boldsymbol{\xi}-i\hat{y}\boldsymbol{B}\cdot\boldsymbol{\xi}}\left\{\left[e^{-i\eta\hat{x}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]}-1\right]+i\eta\hat{x}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]+\frac{1}{2}\eta^2\hat{x}^2\mathcal{G}^2[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\right\}$$

$$+\sum_i k_i^{\rm onl}\frac{\partial}{\partial J_i}P[x,y;\boldsymbol{J}]+\frac{1}{2}\sum_{ij}k_i^{\rm onl}k_j^{\rm onl}\frac{\partial^2}{\partial J_i\partial J_j}P[x,y;\boldsymbol{J}]+\mathcal{O}(N^{-\frac{3}{2}})$$

[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.

We conclude that, at least for the purpose of the expansions relevant to on-line learning, $P[x,y;\boldsymbol{J}]$ is a mean field observable in the sense of (17,18), with the non-trivial contribution of (18) given by

$$\Delta[\boldsymbol{J};\boldsymbol{k}^{\rm onl}]=\frac{1}{p}\left\{\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}-\eta\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]-\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\right.$$
$$\left.+\eta\frac{\partial}{\partial x}\left[\mathcal{G}[x,y]\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\right]-\frac{1}{2}\eta^2\frac{\partial^2}{\partial x^2}\left[\mathcal{G}^2[x,y]\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\right]\right\} \tag{21}$$

Note that $\lim_{N\to\infty}N\Delta[\boldsymbol{J};\boldsymbol{k}^{\rm onl}]=\mathcal{O}(\eta^3/\alpha)$, so that for small learning rates or large training sets this non-trivial term will vanish. Working out the relevant expression for batch learning, on the other hand, gives

$$P[x,y;\boldsymbol{J}+\boldsymbol{k}^{\rm bat}]-P[x,y;\boldsymbol{J}]=\int\frac{d\hat{x}\;d\hat{y}}{(2\pi)^2}e^{i[x\hat{x}+y\hat{y}]}\frac{1}{p}\sum_{\mu=1}^{p}e^{-i\hat{x}\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu-i\hat{y}\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu}$$
$$\times\left\{\left[1-\frac{i\eta\hat{x}}{p}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]+\;\mathcal{O}(N^{-\frac{3}{2}})\right]-1\right\}$$
$$=\sum_i k_i^{\rm bat}\frac{\partial}{\partial J_i}P[x,y;\boldsymbol{J}]+\frac{1}{2}\sum_{ij}k_i^{\rm bat}k_j^{\rm bat}\frac{\partial^2}{\partial J_i\partial J_j}P[x,y;\boldsymbol{J}]+\mathcal{O}(N^{-\frac{3}{2}})$$

Here the term $\Delta[\boldsymbol{J};\boldsymbol{k}^{\rm bat}]$ is absent. In fact also the quadratic contribution $\sum_{ij}k_i^{\rm bat}k_j^{\rm bat}\ldots$ in the above expansion will turn out to be of insignificant order in N. For the purpose of the expansions relevant to batch learning, $P[x,y;\boldsymbol{J}]$ is apparently a simple mean field observable in the sense of (16). This could have been anticipated, since one should ultimately obtain the batch learning equations upon expanding those of on-line learning for small learning rate η, and retaining only the leading order η^1 in this expansion.

2.4 Derivation of Deterministic Dynamical Laws

Having defined our order parameters Q, R and $\{P[x,y]\}$, from this stage onwards the notation $\langle\cdots\rangle_{{\rm QRP};t}$ will be used to denote sub-shell averages defined with respect to these order parameters, at time t. With a modest amount of foresight we define the complementary Kronecker delta $\overline{\delta}_{ab}=1-\delta_{ab}$, and the following key functions:

$$\mathcal{A}[x,y;x',y']=\lim_{N\to\infty}\left\langle\langle\,\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}(\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\,\rangle_{\bar{D}}\rangle_{\bar{D}}\right\rangle_{{\rm QRP};t} \tag{22}$$

$$\mathcal{B}[x,y;x',y']=\lim_{N\to\infty}\left\langle\frac{1}{N}\sum_{i\neq j}\langle\,\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}(\xi_i\xi_j\xi_i'\xi_j')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\,\rangle_{\bar{D}}\rangle_{\bar{D}}\right\rangle_{{\rm QRP};t} \tag{23}$$

$$\mathcal{C}[x,y;x',y';x'',y'']=\lim_{N\to\infty}$$
$$\left\langle\langle\langle\,\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}''}\overline{\delta}_{\boldsymbol{\xi}'\boldsymbol{\xi}''}\frac{(\boldsymbol{\xi}\cdot\boldsymbol{\xi}'')(\boldsymbol{\xi}'\cdot\boldsymbol{\xi}'')}{N}\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\delta[x''-\boldsymbol{J}\cdot\boldsymbol{\xi}'']\delta[y''-\boldsymbol{B}\cdot\boldsymbol{\xi}'']\,\rangle_{\bar{D}}\rangle_{\bar{D}}\rangle_{\bar{D}}\right\rangle_{{\rm QRP};t} \tag{24}$$

We will eventually show in a subsequent section that (23) and (24) are zero. The function (22), on the other hand, will contain all the interesting physics of the learning process, and its calculation will turn out to be our central problem.

In Appendix A we show that for the observables (20) the diffusion matrix elements $G_{\mu\nu}^{\star\star\star}$ in the Fokker-Planck equation (19) vanish for $N\to\infty$. The Fokker-Planck equation (19) now reduces to the Liouville equation $\frac{d}{dt}P_t(\boldsymbol{\Omega})=-\sum_\mu\frac{\partial}{\partial\Omega_\mu}[F_\mu[\boldsymbol{\Omega};t]P_t(\boldsymbol{\Omega})]$, 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:

$$\frac{d}{dt}Q = \lim_{N\to\infty}\left\{2\eta\left\langle\langle(\boldsymbol{J}\cdot\boldsymbol{\xi})\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}\right\rangle_{\mathrm{QRP};t}+\eta^2\left\langle\langle\mathcal{G}^2[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}\right\rangle_{\mathrm{QRP};t}\right\}$$

$$=2\eta\int dxdy\ P[x,y]\ x\ \mathcal{G}[x,y]+\eta^2\int dxdy\ P[x,y]\ \mathcal{G}^2[x,y]$$

$$\frac{d}{dt}R=\lim_{N\to\infty}\eta\left\langle\langle(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}\right\rangle_{\mathrm{QRP};t}=\eta\int dxdy\ P[x,y]\ y\ \mathcal{G}[x,y]$$

These equations are identical to those found in the $\alpha\to\infty$ formalism. The difference is in the function to be substituted for $P[x,y]$, which here is the solution of

$$\frac{\partial}{\partial t}P[x,y]=\lim_{N\to\infty}\left\{-\eta\frac{\partial}{\partial x}\left\langle\langle\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi}',\boldsymbol{B}\cdot\boldsymbol{\xi}'](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}\rangle_{\bar{D}}\right\rangle_{\mathrm{QRP};t}\right.$$

$$+\frac{\eta^2}{2N}\frac{\partial^2}{\partial x^2}\left\langle\langle\mathcal{G}^2[\boldsymbol{J}\cdot\boldsymbol{\xi}',\boldsymbol{B}\cdot\boldsymbol{\xi}'](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')^2\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}\rangle_{\bar{D}}\right\rangle_{\mathrm{QRP};t}$$

$$+\frac{1}{\alpha}\left\langle\langle\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}-\eta\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]-\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}\right.$$

$$\left.\left.+\eta\frac{\partial}{\partial x}\left[\mathcal{G}[x,y]\langle\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}\right]-\frac{1}{2}\eta^2\frac{\partial^2}{\partial x^2}\left[\mathcal{G}^2[x,y]\langle\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\right]\rangle_{\bar{D}}\right\rangle_{\mathrm{QRP};t}\right\}$$

(where we have inserted (21))

$$=\frac{1}{\alpha}\left\{\int dx'\ P[x',y]\delta[x-x'-\eta\mathcal{G}[x',y]]-P[x,y]\right\}$$

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

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

Anticipating the term $\mathcal{B}[\ldots]$ to be zero (as shown in Appendix B) we thus arrive at the following set of coupled deterministic macroscopic equations

$$\frac{d}{dt}Q=2\eta\int dxdy\ P[x,y]\ x\ \mathcal{G}[x,y]+\eta^2\int dxdy\ P[x,y]\ \mathcal{G}^2[x,y] \tag{25}$$

$$\frac{d}{dt}R=\eta\int dxdy\ P[x,y]\ y\ \mathcal{G}[x,y] \tag{26}$$

$$\frac{d}{dt}P[x,y]=\frac{1}{\alpha}\left\{\int dx'\ P[x',y]\delta[x-x'-\eta\mathcal{G}[x',y]]-P[x,y]\right\}$$

$$-\eta\frac{\partial}{\partial x}\int dx'dy'\ \mathcal{A}[x,y;x',y']\ \mathcal{G}[x',y']+\frac{1}{2}\eta^2\int dx'dy'\ P[x',y']\mathcal{G}^2[x',y']\frac{\partial^2}{\partial x^2}P[x,y] \tag{27}$$

Batch Learning

For Q and R one again finds simple equations:

$$\frac{d}{dt}Q = \lim_{N\to\infty}\left\{2\eta\left\langle\langle(\boldsymbol{J}\cdot\boldsymbol{\xi})\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}\right\rangle_{\rm QRP;t} + \frac{\eta^2}{N}\left\langle\sum_i\langle\xi_i\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle^2_{\bar{D}}\right\rangle_{\rm QRP;t}\right\}$$

$$= 2\eta\int dxdy\ P[x,y]\ x\ \mathcal{G}[x;y]$$

$$\frac{d}{dt}R = \lim_{N\to\infty}\eta\left\langle\langle(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}\right\rangle_{\rm QRP;t} = \eta\int dxdy\ P[x,y]\ y\ \mathcal{G}[x;y]$$

Finally we calculate the temporal derivative of the joint field distribution:

$$\frac{\partial}{\partial t}P[x,y] = \lim_{N\to\infty}\left\{-\eta\frac{\partial}{\partial x}\left\langle\langle\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi}',\boldsymbol{B}\cdot\boldsymbol{\xi}'](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}\rangle_{\bar{D}}\right\rangle_{\rm QRP;t}\right.$$

$$\left.+\frac{\eta^2}{2N}\frac{\partial^2}{\partial x^2}\left\langle\langle\langle\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi}',\boldsymbol{B}\cdot\boldsymbol{\xi}']\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi}'',\boldsymbol{B}\cdot\boldsymbol{\xi}'']\rangle(\boldsymbol{\xi}\cdot\boldsymbol{\xi}')(\boldsymbol{\xi}\cdot\boldsymbol{\xi}'')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}\rangle_{\bar{D}}\rangle_{\bar{D}}\right\rangle_{\rm QRP;t}\right\}$$

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

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

Anticipating the term $\mathcal{C}[\ldots]$ to be zero (to be demonstrated in Appendix B) we thus arrive at the following coupled deterministic macroscopic equations:

$$\frac{d}{dt}Q = 2\eta\int dxdy\ P[x,y]\ x\ \mathcal{G}[x;y] \tag{28}$$

$$\frac{d}{dt}R = \eta\int dxdy\ P[x,y]\ y\ \mathcal{G}[x;y] \tag{29}$$

$$\frac{d}{dt}P[x,y] = -\frac{\eta}{\alpha}\frac{\partial}{\partial x}\left[\mathcal{G}[x,y]P[x,y]\right] - \eta\frac{\partial}{\partial x}\int dx'dy'\ \mathcal{A}[x,y;x',y']\ \mathcal{G}[x',y'] \tag{30}$$

The difference between the macroscopic equations for batch and on-line learning is merely the presence (on-line) or absence (batch) of those terms which are not linear in the learning rate η (i.e. of order η^2 or higher).

2.5 Closure of Macroscopic Dynamical Laws

The complexity of the problem is fully concentrated in the Green's function $\mathcal{A}[x,y;x',y']$ defined in (22). Our macroscopic laws are exact for $N\to\infty$ but not yet closed due to the appearance of the microscopic probability density $p_t(\boldsymbol{J})$ in the sub-shell average of (22). We now close our macroscopic laws by making, for $N\to\infty$, the two key assumptions underlying dynamical replica theories:

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

Assumption 1 implies that all microscopic probability variations within the $\{Q,R,P\}$ sub-shells 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 (22): microscopic probability equipartitioning in the $\{Q,R,P\}$-subshells of the ensemble, or

$$p_t(\boldsymbol{J}) \quad \to \quad w(\boldsymbol{J}) \;\sim\; \delta[Q-Q[\boldsymbol{J}]]\;\delta[R-R[\boldsymbol{J}]]\;\prod_{xy}\delta[P[x,y]-P[x,y;\boldsymbol{J}]] \tag{31}$$

This new microscopic distribution $w(\boldsymbol{J})$ depends on time via the order parameters $\{Q,R,P\}$. Note that (31) leads to exact macroscopic laws if our observables $\{Q,R,P\}$ for $N\to\infty$ 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 using the path integral formalism (see e.g. [10]). We write averages over all training sets $\tilde{D}\subseteq\{-1,1\}^N$, with $|\tilde{D}|=p$, as $\langle\ldots\rangle_{\Xi}$. Our assumptions result in the closure of the two sets (25,26,27) and (28,29,30), since now the function $\mathcal{A}[x,y;x',y']$ is expressed fully in terms of $\{Q,R,P\}$:

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

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

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

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

$$\tilde{w}(\boldsymbol{\sigma})\;\sim\;\delta\left[N-\boldsymbol{\sigma}^2\right]\delta\left[NRQ^{-\frac{1}{2}}-\boldsymbol{\tau}\cdot\boldsymbol{\sigma}\right]\prod_{xy}\delta\left[P[x,y]-P[x,y;(Q/N)^{\frac{1}{2}}\boldsymbol{\sigma}]\right] \tag{32}$$

We thus arrive at

$$\mathcal{A}[x,y;x',y'] = \lim_{\substack{N\to\infty\\ n\to 0}}\int\prod_{\alpha=1}^n\tilde{w}(\boldsymbol{\sigma}^\alpha)d\boldsymbol{\sigma}^\alpha$$
$$\left\langle\;\left\langle\!\!\left\langle(\boldsymbol{\xi}'\cdot\boldsymbol{\xi})\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}\,\delta\left[x-\frac{\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]\delta\left[y-\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]\delta\left[x'-\frac{\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right]\delta\left[y'-\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right]\right\rangle_{\bar{D}}\right\rangle_{\bar{D}}\;\right\rangle_{\Xi} \tag{33}$$

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):

$$P[x,y] = \lim_{\substack{N\to\infty\\ n\to 0}}\int\prod_{\alpha=1}^n\tilde{w}(\boldsymbol{\sigma}^\alpha)d\boldsymbol{\sigma}^\alpha\left\langle\left\langle\delta\left[x-\frac{\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]\delta\left[y-\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]\right\rangle_{\bar{D}}\right\rangle_{\Xi} \tag{34}$$

Finally we will have to demonstrate that the two functions $\mathcal{B}[\ldots]$ and $\mathcal{C}[\ldots]$, as defined in (23,24), do indeed vanish self-consistently, as claimed. To achieve this we again express them in replica form:

$$\mathcal{B}[x,y;x',y'] = \lim_{\substack{N\to\infty\\ n\to 0}}\int\prod_{\alpha=1}^n\tilde{w}(\boldsymbol{\sigma}^\alpha)d\boldsymbol{\sigma}^\alpha$$

$$\left\langle \left\langle\!\!\left\langle \overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'} \left[\frac{1}{N}\sum_{i\neq j}\xi_i\xi_j\xi_i'\xi_j'\right] \delta\left[x-\frac{\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}}{\sqrt{N}}\right] \delta\left[y-\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right] \delta\left[x'-\frac{\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \delta\left[y'-\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \right\rangle_{\bar{D}} \right\rangle_{\bar{D}} \right\rangle_{\boldsymbol{\Xi}} \tag{35}$$

and

$$\mathcal{C}[x,y;x',y';x'',y''] = \lim_{\substack{N\to\infty\\ n\to 0}} \int\prod_{\alpha=1}^{n}\tilde{w}(\boldsymbol{\sigma}^\alpha)d\boldsymbol{\sigma}^\alpha \left\langle \left\langle\!\!\left\langle \overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}''}\overline{\delta}_{\boldsymbol{\xi}'\boldsymbol{\xi}''}\frac{(\boldsymbol{\xi}\cdot\boldsymbol{\xi}'')(\boldsymbol{\xi}'\cdot\boldsymbol{\xi}'')}{N}\delta\left[x-\frac{\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}}{\sqrt{N}}\right] \delta\left[y-\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]\right.\right.\right.$$
$$\left.\left.\left. \times\,\delta\left[x'-\frac{\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \delta\left[y'-\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \delta\left[x''-\frac{\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}''}{\sqrt{N}}\right] \delta\left[y''-\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}''}{\sqrt{N}}\right] \right\rangle_{\bar{D}} \right\rangle_{\bar{D}} \right\rangle_{\bar{D}} \right\rangle_{\boldsymbol{\Xi}} \tag{36}$$

At this stage the physics is over, what remains is to perform the summations and integrations in (33,34,35,36) in the limit $N\to\infty$. Full details of this exercise are given in Appendix B, where we show that (35) and (36) are indeed zero, and where we derive, in replica symmetric ansatz, an expression for the Green's function (33). It turns out that to calculate this Green's function $\mathcal{A}[\ldots]$ one has to solve two coupled saddle-point equations at each time-step, one scalar equation relating to a spin-glass order parameter q, and one functional saddle-point equation relating to an effective single-spin measure.

3 Summary of the Theory and Connection with $\alpha\to\infty$ Formalism

In this section we summarize the results obtained so far (including the replica calculation in Appendix B) in a compact way, and we show that our general theory has the satisfactory property that it incorporates the standard formalism developed for infinite training sets (with Gaussian joint field distributions $P[x,y]$ at any time) as a special case, recovered in the limit $\alpha\to\infty$. In addition we provide a proof of the uniqueness of the RS functional saddle-point equation and show that it can be found as the fixed-point of an iterative map.

3.1 Summary of the Theory

Dynamic Equations for Observables

Our observables are $Q=\boldsymbol{J}^2$, $R=\boldsymbol{J}\cdot\boldsymbol{B}$, and the joint distribution of student and teacher fields $P[x,y]=\langle\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar{D}}$. For $N\to\infty$ these quantities obey closed, deterministic, and self-averaging macroscopic dynamic equations. One always has $P[x,y]=P[x|y]P[y]$ with $P[y]=(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}y^2}$. We define $\langle f[x,y]\rangle=\int dxDy\; P[x|y]f[x,y]$, with the familiar short-hand $Dy=(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}y^2}dy$, and the following four averages (the function $\Phi[x,y]$ will be given below):

$$U=\langle\Phi[x,y]\mathcal{G}[x,y]\rangle \qquad V=\langle x\mathcal{G}[x,y]\rangle \qquad W=\langle y\mathcal{G}[x,y]\rangle \qquad Z=\langle\mathcal{G}^2[x,y]\rangle \tag{37}$$

For on-line learning our macroscopic laws are

$$\frac{d}{dt}Q=2\eta V+\eta^2 Z \qquad\qquad \frac{d}{dt}R=\eta W \tag{38}$$

$$\frac{d}{dt}P[x|y]=\frac{1}{\alpha}\int dx'P[x'|y]\left[\delta[x-x'-\eta\mathcal{G}[x',y]]-\delta[x-x']\right]-\eta\frac{\partial}{\partial x}\left\{P[x|y]\left[U(x-Ry)+Wy\right]\right\}$$
$$+\frac{1}{2}\eta^2 Z\frac{\partial^2}{\partial x^2}P[x|y]-\eta\left[V-RW-(Q-R^2)U\right]\frac{\partial}{\partial x}\left\{P[x|y]\Phi[x,y]\right\} \tag{39}$$

For batch learning one has:

$$\frac{d}{dt}Q=2\eta V \qquad\qquad \frac{d}{dt}R=\eta W \tag{40}$$
$$\frac{d}{dt}P[x|y]=-\frac{\eta}{\alpha}\frac{\partial}{\partial x}\left[P[x|y]\mathcal{G}[x,y]\right]-\eta\frac{\partial}{\partial x}\left\{P[x|y]\left[U(x-Ry)+Wy\right]\right\}$$

$$-\eta\,[V-RW-(Q-R^2)U]\,\frac{\partial}{\partial x}\left\{P[x|y]\Phi[x,y]\right\} \tag{41}$$

Note that the batch equations follow from the on-line ones by retaining only terms which are linear in the learning rate. From the solution of the above equations follow, in turn, the training- and generalization errors:

$$E_{\rm t}=\langle\theta[-xy]\rangle \qquad\qquad E_{\rm g}=\frac{1}{\pi}\arccos[R/\sqrt{Q}] \tag{42}$$

We note, finally, that the first conditional moment $\overline{x}(y)=\int dx\; xP[x|y]$ of $P[x|y]$ of the joint field distribution obeys a simple equation, which is obtained from (39) and (41) upon multiplication by x, followed by integration over x:

$$\frac{d}{dt}[\overline{x}(y)-Ry]=\frac{\eta}{\alpha}\int dx\; P[x|y]\mathcal{G}[x,y]+\eta U[\overline{x}(y)-Ry] \tag{43}$$

where we have also used the built-in property $\int dx\; P[x|y]\Phi[x,y]=0$ for all y.

Saddle-Point Equations and the function Φ

The function $\Phi[x,y]$ appearing in the above equations (generated by the Green's function $\mathcal{A}[\ldots]$) is expressed in terms of auxiliary order parameters. These come about in the replica calculation of Appendix B, where the order parameters are defined through Dirac δ functions in their integral representation. The first auxiliary order parameter is a spin-glass type order parameter $q=\langle\langle\boldsymbol{J}\rangle^2\rangle_{\bar{D}}/Q$, with $R^2/Q\leq q\leq 1$. The second, defined similarly for the joint probability $P[x,y]$ is the function $\chi[x,y]$ (for details see Appendix B). The latter is not necessarily normalised and in what follows it is useful to consider the effective measure $M[x,y]$ which is related to $\chi[x,y]$ through a simple transformation (equation (146)). The measure $M[x,y]$ is non-negative and can be always normalized such that $\int dx\; M[x,y]=1$ for all $y\in\Re$, as emphasized in our notation by writing $M[x,y]\to M[x|y]$. The auxiliary order parameters are calculated at each time-step by solving the following two coupled saddle-point equations:

$$\langle(x-Ry)^2\rangle+(qQ-R^2)(1-\frac{1}{\alpha})=\left[\frac{1+q-2R^2/Q}{1-q}\right]\int DyDz\;\left[\langle x^2\rangle_\star-\langle x\rangle_\star^2\right] \tag{44}$$

$$P[X|y]=\int Dz\;\langle\delta[X-x]\rangle_\star \tag{45}$$

in which

$$\langle f[x,y,z]\rangle_\star=\frac{\int dx\; M[x|y]e^{Bxz}f[x,y,z]}{\int dx\; M[x|y]e^{Bxz}} \qquad\qquad B=\frac{\sqrt{qQ-R^2}}{Q(1-q)} \tag{46}$$

After q and $M[x|y]$ have been determined, the key function $\Phi[x,y]$ in (37,39,41) is calculated as

$$\Phi[X,y]=\left\{Q(1-q)P[X|y]\right\}^{-1}\int Dz\langle X-x\rangle_\star\langle\delta[X-x]\rangle_\star \tag{47}$$

or, equivalently:

$$\Phi[X,y]=\left\{\sqrt{qQ-R^2}P[X|y]\right\}^{-1}\int Dz\; z\;\langle\delta[X-x]\rangle_\star \tag{48}$$

Finding a saddle-point problem for an order parameter function, rather than a finite number of scalar order parameters, introduces the possibility of a proliferation of saddle-points. In the next section we will show that this does not happen: the solution of the functional saddle-point problem is unique, and can even by found iteratively by executing a specific non-linear mapping.

3.2 Uniqueness and Iterative Calculation of the Functional Saddle-Point

The uniqueness proof is more easily set up in terms of the original order parameter function $\chi[x,y]$, rather than the new (normalised) measure $M[x|y]$ (see Appendix B). For a given state $\{Q,R,P\}$ and a given value for $q\in[R^2/Q,1]$ we have to find the functional saddle-points of the functional $\Psi[\chi]$, defined as:

$$\Psi[\chi]=\alpha\int DyDz\ \log\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\alpha^{-1}\chi[x,y]}-\int DydxP[x|y]\chi[x,y] \tag{49}$$

Our proof will carry the existence of the various integrals as an implicit condition for validity. To reduce notational ballast we define

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

Note: $w(x,y,z)=M[x|y]e^{Bxz}/\int dx'\ M[x'|y]e^{Bx'z}$. The function $w(u,v,z)$ obeys

$$\frac{\delta w(u,v,z)}{\delta\chi[u',v']}=\alpha^{-1}\delta[v-v']\left[\delta[u-u']w(u,v,z)-w(u,v,z)w(u',v,z)\right]$$

The functional saddle-point equation is obtained by requiring the first functional derivative of $\Psi[\chi]$ with respect to $\chi[u,v]$ to be zero for all $u,v\in\Re$, where

$$\left.\frac{\delta\Psi}{\delta\chi[u,v]}\right|_\chi=\frac{e^{-\frac{1}{2}v^2}}{\sqrt{2\pi}}\left\{\int Dz\ w(u,v,z)-P[u|v]\right\} \tag{50}$$

Clearly, if the function $\chi[x,y]$ is a saddle-point, then also the function $\chi[x,y]+\rho(y)$ for any $\rho(y)$. This degree of freedom is irrelevant because such terms $\rho(y)$ will drop out of the measure $\langle\ldots\rangle_\star$. Furthermore, one immediately verifies that transformations of the form $\chi[x,y]\to\chi[x,y]+\rho(y)$ leave the functional $\Psi[\ldots]$ (49) invariant. Next we calculate the Hessian (or curvature) operator $H[u,v;u',v';\chi]$, using (50):

$$H[u,v;u',v';\chi]=\left.\frac{\delta^2\Psi}{\delta\chi[u,v]\delta\chi[u',v']}\right|_\chi=\frac{e^{-\frac{1}{2}v^2}}{\sqrt{2\pi}}\int Dz\ \frac{\delta w(u,v,z)}{\delta\chi[u',v']}$$

$$=\delta[v-v']\frac{e^{-\frac{1}{2}v^2}}{\alpha\sqrt{2\pi}}\int Dz\left[\delta[u-u']w(u,v,z)-w(u,v,z)w(u',v,z)\right] \tag{51}$$

$H[u,v;u',v';\chi]$ is non-negative definite for each χ, and thus the functional Ψ is convex, since for any function $\phi[u,v]$ for which the relevant integrals exist we find

$$\int dudvdu'dv'\ \phi[u,v]H[u,v;u',v';\chi]\phi[u',v']=\frac{1}{\alpha}\int DvDz\ \left[\langle\phi^2[u,v]\rangle_\star-\langle\phi[u,v]\rangle_\star^2\right]\geq 0$$

The kernel of $H[u,v;u',v';\chi]$, for a given 'point' χ in χ-space, is determined by requiring equality in the above inequality, i.e.

$$\text{for each}\ \ v,z\in\Re:\qquad\qquad \langle[\phi[u,v]-\langle\phi[u,v]\rangle_\star]^2\rangle_\star=0\quad\ \text{so}\quad\ \frac{\partial}{\partial u}\phi[u,v]=0$$

For each χ the kernel of the second functional derivative $H[x,y;x',y';\chi]$ thus consists of the set of all (integrable) functions $\phi[x,y]$ which depend on y only.

We now find that, if $\chi_0[x,y]$ and $\chi_1[x,y]$ are both functional saddle-points of $\Psi[\chi]$, then $\chi_1[x,y]-\chi_0[x,y]=\rho(y)$ for some function $\rho(y)$. In other words: apart from the aforementioned irrelevant degree of freedom, the solution of the functional saddle-point equation (45) is unique. To show this, consider two functions $\chi_0[x,y]$

and $\chi_1[x,y]$ which are both functional saddle-points of Ψ, i.e. corresponding to solutions of (45). Define a path $\{\chi_t\}$ through χ-space, connecting these two functions:

$$\chi_t[x,y] = \chi_0[x,y] + t\left\{\chi_1[x,y] - \chi_0[x,y]\right\}, \qquad t \in [0,1]$$

Integration along this path will bring us from χ_0 to χ_1. Thus for any functional $L[\chi]$ one has

$$L[\chi_1] - L[\chi_0] = \int_{\chi_0}^{\chi_1} dL[\chi] = \int dudv \int_{\chi_0}^{\chi_1} d\chi[u,v] \frac{\delta L}{\delta \chi[u,v]}$$

$$= \int dudv\, [\chi_1[u,v] - \chi_0[u,v]] \int_0^1 dt\, \frac{\delta L}{\delta \chi[u,v]}\bigg|_{\chi_t}$$

For the functional $L[\chi]$ we now choose a functional first derivative of $\Psi[\chi]$, i.e. $L[\chi] = \delta\Psi/\delta\chi[x,y]$ for some $x, y \in \Re$. Since both χ_0 and χ_1 are saddle-points one finds $L[\chi_0] = L[\chi_1] = 0$. Thus

$$\int dudv\, [\chi_1[u,v] - \chi_0[u,v]] \int_0^1 dt\, \frac{\delta^2 \Psi}{\delta\chi[u,v]\delta\chi[x,y]}\bigg|_{\chi_t} = 0$$

Multiply both sides by $\chi_1[x,y] - \chi_0[x,y]$ and integrate the result over $x, y \in \Re$:

$$\int_0^1 dt \int dudvdxdy\, [\chi_1[u,v] - \chi_0[u,v]]\, H[u,v;x,y;\chi_t]\, [\chi_1[x,y] - \chi_0[x,y]] = 0$$

One concludes (since the Hessian is a symmetric non-negative operator):

$$\text{for all } \; t \in [0,1],\; u,v \in \Re: \qquad \int dxdy\; H[u,v;x,y;\chi_t]\, [\chi_1[x,y] - \chi_0[x,y]] = 0$$

The function $\chi_1[x,y] - \chi_0[x,y]$ is in the kernel of $H|_{\chi_t}$ for any $t \in [0,1]$. The kernel of H was already determined to be the set of all integrable functions which depend on y only, whatever the point χ where one chooses to evaluate H. Hence $\chi_1[x,y] - \chi_0[x,y] = \rho(y)$ for some function $\rho(y)$. Finally, the remaining freedom in choosing a function ρ is eliminated by our normalisation $\int dx\; M[x|y] = 1$ (for each y), so that the solution $M[x|y]$ is indeed truly unique.

Next we will show how for any given value of the scalar order parameter q and the observables $\{Q, R, P\}$ (and thus of B), for which the relevant integrals exist, the unique solution $M[x|y]$ of the functional saddle-point equation (45) can be constructed as the stable fixed-point of the following functional map:

$$\text{for each } y \in \Re: \qquad M_{\ell+1}[x|y] = \frac{P[x|y] \left\{ \int Dz \left[\int dx'\; e^{Bz(x'-x)} M_\ell[x'|y] \right]^{-1} \right\}^{-1}}{\int du\; P[u|y] \left\{ \int Dz \left[\int dx'\; e^{Bz(x'-u)} M_\ell[x'|y] \right]^{-1} \right\}^{-1}} \tag{52}$$

Clearly all fixed-points of this map correspond to normalised solutions $M[x|y]$ of a functional saddle-point equation (45), of which there can be only one. Thus we only need to verify the convergence of (52), which can be done most efficiently using an appropriate Lyapunov functional. Note that the functional (49) can be written as

$$\Psi[M|y] = \alpha \int Dy\; \tilde{\Psi}[M|y] \;+\; \text{terms independent of } M[\ldots]$$

with

$$\tilde{\Psi}[M|y] = \int Dz\; \log \int dx\; M[x|y] e^{Bxz} - \int dx P[x|y] \log M[x|y] \tag{53}$$

For any given $y \in \Re$ we will show (53) to be a Lyapunov functional for the mapping (52), i.e. $\tilde{\Psi}[M|y]$ is bounded from below and monotonically increasing during the iteration of (52) with stationarity obtained only

when $M[\ldots]$ is the (unique) fixed-point of (52). First we prove that a lower bound for $\tilde{\Psi}$ is given by the entropy of the conditional distribution $P[x|y]$:

$$\text{for any } M[\ldots] \text{ and any } y \in \Re: \qquad \tilde{\Psi}[M|y] \;\geq\; -\int dx\; P[x|y] \log P[x|y] \tag{54}$$

The proof is elementary (using Jenssen's inequality):

$$\tilde{\Psi}[M|y] = \int Dz\; \log\left\{\int dx\; P[x|y] e^{Bxz+\log M[x|y]-\log P[x|y]}\right\} - \int dx\; P[x|y]\log M[x|y]$$

$$\geq \int Dz \int dx\; P[x|y]\left\{Bxz + \log M[x|y] - \log P[x|y]\right\} - \int dx\; P[x|y]\log M[x|y]$$

$$= -\int dx\; P[x|y]\log P[x|y]$$

Secondly we show that (53) indeed decreases monotonically under (52) until the fixed-point of (52) is reached. To do so we introduce the short-hand notations $\lambda_\ell(x,y,z) = Bxz + \log M_\ell[x|y] - \log P[x|y]$, $\langle f[x]\rangle = \int dx\; P[x|y] f[x]$, and

$$v_\ell(x,y) = \left\{\int Dz\; e^{\lambda_\ell(x,y,z)} \langle e^{\lambda_\ell(x',y,z)}\rangle^{-1}\right\}^{-1}$$

The iterative map can now be written as

$$M_{\ell+1}[x|y] = \frac{M_\ell[x|y] v_\ell(x,y)}{\int du\; M_\ell[u|y] v_\ell(u,y)}$$

This gives for the change in $\tilde{\Psi}[\ldots]$ during one iteration of the mapping, again with Jenssen's inequality:

$$\tilde{\Psi}[M_{\ell+1}|y] - \tilde{\Psi}[M_\ell|y] = \int Dz\; \log\left\{\frac{\int dx\; M_{\ell+1}[x|y] e^{Bxz}}{\int dx\; M_\ell[x|y] e^{Bxz}}\right\} - \int dx P[x|y] \log\left\{\frac{M_{\ell+1}[x|y]}{M_\ell[x|y]}\right\}$$

$$= \int Dz \left\{\log \frac{\langle e^{\lambda_\ell(x,y,z)} v_\ell(x,y)\rangle}{\langle e^{\lambda_\ell(x,y,z)}\rangle}\right\} - \langle \log v_\ell(x,y)\rangle$$

$$\leq\; \log\left\{\langle v_\ell(x,y) \int Dz \left[e^{\lambda_\ell(x,y,z)} \langle e^{\lambda_\ell(x',y,z)}\rangle^{-1}\right]\rangle\right\} - \langle \log v_\ell(x,y)\rangle$$

$$= -\langle \log v_\ell(x,y)\rangle = \langle \log \int Dz\; e^{\lambda_\ell(x,y,z)} \langle e^{\lambda_\ell(x',y,z)}\rangle^{-1}\rangle$$

$$\leq\; \log \int Dz\; \langle e^{\lambda_\ell(x,y,z)}\rangle \langle e^{\lambda_\ell(x',y,z)}\rangle^{-1} = 0$$

Finally we round off our argument by inspecting the implications of having strict equality in the above inequality. Equality can only occur if at both instances where Jenssen's inequality was used in replacements of the form $\langle \log(X)\rangle \leq \log\langle X\rangle$ the relevant stochastic variable X was a constant. In our problem this gives the two conditions

$$\frac{\partial}{\partial z} \frac{\langle e^{\lambda_\ell(x,y,z)} v_\ell(x,y)\rangle}{\langle e^{\lambda_\ell(x,y,z)}\rangle} = 0, \qquad \frac{\partial}{\partial x} v_\ell(x,y) = 0$$

If the second condition is met, the first immediately follows. Working out the second condition gives, in combination with the property that $P[x|y]$ is normalised:

$$\int Dz \frac{M_\ell[x|y] e^{Bxz}}{\int dx'\; M_\ell[x'|y] e^{Bx'z}} = P[x|y]$$

Thus we have confirmed that $\tilde{\Psi}[M_{\ell+1}|y] = \tilde{\Psi}[M_\ell|y]$ if and only if $M_\ell[\ldots]$ is the (unique) fixed-point of (52).

As a consequence of the above we may now write the normalised solution of our functional saddle-point equation (45) in terms of repeated execution of the mapping (52) following an an in principle arbitrary initialisation:

$$\text{for all } y \in \Re: \qquad M[x|y] = \lim_{\ell\to\infty} M_\ell[x|y], \qquad M_0[x|y] = P[x|y]$$

This property simplifies the numerical solution of our equations drastically.

3.3 Fourier Representation and Conditionally-Gaussian Solutions

There are two potential advantages of rewriting our equations in Fourier representation. Firstly, after a Fourier transform the functional saddle-point equation (45) will acquire a much simpler form. Secondly, in those cases where we expect $P[x|y]$ to be of a Gaussian shape in x this would simplify solution of the diffusion equations (39,41). Clearly, $P[x,y]$ being Gaussian in (x,y) is not equivalent to $P[x|y]$ being Gaussian in x only. The former requires

$$\frac{\partial^2}{\partial y^2}\int dx\; xP[x|y] = \frac{\partial}{\partial y}\left\{\int dx\; x^2P[x|y] - \left[\int dx\; xP[x|y]\right]^2\right\} = 0,$$

which only will turn out to happen for $\alpha\to\infty$. A Gaussian $P[x|y]$ with moments which depend on y in a non-trivial way, on the other hand, is found to occur also for $\alpha<\infty$, provided we consider simple learning rules and small η. To avoid ambiguity we will call solutions of the latter type 'conditionally-Gaussian'.

We introduce the Fourier transforms

$$\hat{P}[k|y] = \int dx\; e^{-ikx}P[x|y] \qquad\qquad \hat{M}[k|y] = \int dx\; e^{-ikx}M[x|y] \tag{55}$$

The transformed functional saddle-point equation thereby acquires a very simple form

$$\hat{P}[k|y] = \int Dz\; \frac{\hat{M}[k+iBz|y]}{\hat{M}[iBz|y]} \tag{56}$$

Note that, in contrast to the original equation (45), the transformed equation (56) need not have a unique solution (it could allow for solutions corresponding to non-integrable functions in the original problem). Consider, for instance, the transformation

$$\hat{M}[k|y] \;\to\; \dot{M}[k|y] = \frac{e^{\frac{1}{2}k^2/B^2}}{\hat{M}[-k|y]}$$

with the property (verified by a simple transformation of variables):

$$\int Dz\; \frac{\dot{M}[k+iBz|y]}{\dot{M}[iBz|y]} = \int_{ik/B-\infty}^{ik/B+\infty} Dz\; \frac{\hat{M}[k+iBz|y]}{\hat{M}[iBz|y]}$$

If $\hat{M}[k]$, which by definition cannot have poles, is sufficiently well behaved, a simple deformation of the integration path (via contour integration) leads to the statement that if $\hat{M}[k|y]$ is a solution of (56), then so is $\dot{M}[k|y]$.

Transformation of the dynamical on-line equation (39) for $P[x|y]$ (from the which the batch equation (41) can be obtained by expansion in η) gives:

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

$$+\;\eta kU\frac{\partial}{\partial k}\log\hat{P}[k|y] - \frac{1}{2}\eta^2k^2Z \;-\; i\eta k\left[\frac{V-RW-(Q-R^2)U}{\sqrt{qQ-R^2}\hat{P}[k|y]}\right]\int Dz\; z\;\frac{\hat{M}[k+iBz|y]}{\hat{M}[iBz]} \tag{57}$$

We now determine the conditions for equation (57) to have conditionally-Gaussian solutions. If $P[x|y]$ is Gaussian in x we can solve the functional saddle-point equation (45) (whose solution is unique), and find the resulting pair of measures

$$P[x|y] = \frac{e^{-\frac{1}{2}[x-\overline{x}(y)]^2/\Delta^2(y)}}{\Delta(y)\sqrt{2\pi}} \qquad\qquad M[x|y] = \frac{e^{-\frac{1}{2}[x-\overline{x}(y)]^2/\sigma^2(y)}}{\sigma(y)\sqrt{2\pi}} \tag{58}$$

$$\Delta^2(y) = \sigma^2(y) + B^2\sigma^4(y) \tag{59}$$

with their Fourier transforms $\hat{P}[k|y] = \exp\left[-ik\overline{x}(y) - \frac{1}{2}k^2\Delta^2(y)\right]$ and $\hat{M}[k|y] = \exp\left[-ik\overline{x}(y) - \frac{1}{2}k^2\sigma^2(y)\right]$. Insertion of these expressions as an Ansatz into (57), using the identity

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

and performing some simple manipulations, gives the following simplified equation:

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

$$-\frac{1}{2}k^2\left\{\eta^2 Z + 2\eta U\Delta^2(y)\ + 2\eta\sigma^2(y)\left[\frac{V - RW - (Q-R^2)U}{Q(1-q)}\right]\right\} \tag{60}$$

From this it follows that conditionally-Gaussian solutions can occur in two situations only:

$$\alpha\to\infty \qquad \text{or} \qquad \frac{\partial^3}{\partial k^3}\int\frac{du}{\sqrt{2\pi}}\ e^{-\frac{1}{2}[u-ik\Delta(y)]^2 - ik\eta\mathcal{G}[\overline{x}(y)+u\Delta(y),y]} = 0 \tag{61}$$

The first case corresponds to the familiar theory of infinite 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 of (61) we retain only the term linear in η) or with on-line execution for small η (retaining in (61) only η and η^2 terms). The latter cases will be dealt with in more detail later.

3.4 Link with the Formalism for Complete Training Sets

The very least we should require of our theory is that it reduces to the simple (Q,R) formalism of complete training sets [2, 3] in the limit $\alpha\to\infty$. Here we will show that this indeed happens. In the previous section we have seen that for $\alpha\to\infty$ our driven diffusion equation for the conditional distribution $P[x|y]$ has conditionally-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 $\Phi[x,y]$ (47) 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 dynamical 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 equations reduce to:

$$\frac{d}{dt}Q = \begin{cases} 2\eta\langle x\mathcal{G}[x,y]\rangle + \eta^2\langle\mathcal{G}^2[x,y]\rangle & (\text{on-line}) \\ 2\eta\langle x\mathcal{G}[x,y]\rangle & (\text{batch}) \end{cases} \qquad \frac{d}{dt}R = \eta\langle y\mathcal{G}[x,y]\rangle \tag{62}$$

$$\frac{1}{\eta}\frac{d}{dt}\left[\overline{x}(y) - Ry\right] = [\overline{x}(y) - Ry]\langle\Phi[x',y']\mathcal{G}[x',y']\rangle \tag{63}$$

$$\frac{1}{2\eta}\frac{d}{dt}\left[\Delta^2(y)-Q+R^2\right] = \langle (x'-Ry')\mathcal{G}[x',y']\rangle\left[\frac{\sigma^2(y)}{Q(1-q)}-1\right] + \langle \Phi[x',y']\mathcal{G}[x',y']\rangle\left[\Delta^2(y)-\frac{Q-R^2}{Q(1-q)\sigma^2(y)}\right] \quad (64)$$

with one remaining saddle-point equation to determine q, obtained upon working out (44) for conditionally-Gaussian solutions:

$$\int Dy\,\{[\overline{x}(y)-Ry]^2+\Delta^2(y)\}+qQ-R^2 = \left[2\frac{qQ-R^2}{Q(1-q)}+1\right]\int Dy\;\sigma^2(y) \quad (65)$$

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

$$P[x|y] = \frac{e^{-\frac{1}{2}[x-Ry]^2/(Q-R^2)}}{\sqrt{2\pi(Q-R^2)}}, \quad (66)$$

Insertion into the dynamical equations shows that (63) is now immediately satisfied, that (64) reduces to $\sigma^2(y)=Q(1-q)$, and that as a result the saddle-point equation (65) is automatically satisfied. Since (66) is parametrized by Q and R only, this leaves us with the closed equations

$$\frac{d}{dt}Q = \begin{cases} 2\eta\langle x\mathcal{G}[x,y]\rangle+\eta^2\langle\mathcal{G}^2[x,y]\rangle & (\text{on-line}) \\ 2\eta\langle x\mathcal{G}[x,y]\rangle & (\text{batch}) \end{cases} \qquad \frac{d}{dt}R = \eta\langle y\mathcal{G}[x,y]\rangle \quad (67)$$

These are the equations found in e.g. [2, 3]. From our general theory for restricted training sets we thus indeed recover in the limit $\alpha\to\infty$ the standard formalism (66,67) describing learning with complete training sets, as claimed.

4 Benchmark Tests: Hebbian Learning

In the special case of the Hebb rule, $\mathcal{G}[x,y]=\text{sgn}[y]$, 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:

$$P_0[x|y] = \frac{e^{-\frac{1}{2}(x-R_0y)^2/(Q_0-R_0^2)}}{\sqrt{2\pi(Q_0-R_0^2)}} \quad (68)$$

Careful averaging of the exact expressions for our observables over all 'paths' $\{\boldsymbol{\xi}(0),\boldsymbol{\xi}(1),\ldots\}$ taken by the question/example 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 [12]. For on-line Hebbian learning one ends up with:

$$Q = Q_0+2\eta tR_0\sqrt{\frac{2}{\pi}}+\eta^2t+\eta^2t^2\left[\frac{1}{\alpha}+\frac{2}{\pi}\right] \qquad R = R_0+\eta t\sqrt{\frac{2}{\pi}} \quad (69)$$

$$P[x|y] = \int\frac{d\hat{x}}{2\pi}\;e^{-\frac{1}{2}\hat{x}^2[Q-R^2]+i\hat{x}[x-Ry]+\frac{t}{\alpha}[e^{-i\eta\hat{x}\;\text{sgn}[y]}-1]} \quad (70)$$

For batch learning a similar calculation[3] gives:

$$Q = Q_0+2\eta tR_0\sqrt{\frac{2}{\pi}}+\eta^2t^2\left[\frac{1}{\alpha}+\frac{2}{\pi}\right] \qquad R = R_0+\eta t\sqrt{\frac{2}{\pi}} \quad (71)$$

$$P[x|y] = \frac{e^{-\frac{1}{2}[x-Ry-(\eta t/\alpha)\;\text{sgn}[y]]^2/(Q-R^2)}}{\sqrt{2\pi(Q-R^2)}} \quad (72)$$

[3]Note that in [12] only the on-line calculation was carried out; the batch calculation can be done along the same lines.

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

$$\int dx\ xP[x|y] = Ry + \frac{\eta t}{\alpha}\ \mathrm{sgn}[y] \tag{73}$$

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 $E_{\rm t} = \int DydxP[x|y]\theta[-xy]$. We thus have exact expressions for both the generalization error and the training error at any time and for any α. The asymptotic values, for both batch and on-line Hebbian learning, are given by

$$\lim_{t\to\infty} E_{\rm g} = \frac{1}{\pi}\arccos\left[\frac{1}{\sqrt{1+\pi/2\alpha}}\right] \tag{74}$$

$$\lim_{t\to\infty} E_{\rm t} = \frac{1}{2} - \frac{1}{2}\int Dy\ \mathrm{erf}\left[|y|\sqrt{\frac{\alpha}{\pi}} + \frac{1}{\sqrt{2\alpha}}\right] \tag{75}$$

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 the one investigated in the present paper.

4.1 Batch Hebbian Learning

We 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 (68), we immediate obtain $V_0 = R_0\sqrt{2/\pi}$. 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

$$Q = Q_0 + 2\eta t R_0\sqrt{\frac{2}{\pi}} + \eta^2 t^2\left[\frac{1}{\alpha} + \frac{2}{\pi}\right] \qquad\qquad R = R_0 + \eta t\sqrt{\frac{2}{\pi}} \tag{76}$$

Furthermore, it turns out that the above diffusion equation for $P[x|y]$ meets the requirements for having conditionally-Gaussian solutions, i.e.

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

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

$$\overline{x}(y) = Ry + \frac{\eta t}{\alpha}\ \mathrm{sgn}(y) \qquad\qquad \frac{d}{dt}\Delta^2(y) = \frac{2\eta^2 t\sigma^2(y)}{\alpha Q(1-q)} \qquad\qquad \Delta^2(y) = \sigma^2(y) + B^2\sigma^4(y)$$

The spin-glass order parameter q is to be solved from the remaining scalar saddle-point equation (44). With help of identities like $\langle x\rangle_\star = \overline{x}(y) + zB\sigma^2(y)$, which only hold for conditionally-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

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

$$P[x|y] = \frac{e^{-\frac{1}{2}[x-Ry-(\eta t/\alpha)\ \mathrm{sgn}(y)]^2/(Q-R^2)}}{\sqrt{2\pi(Q-R^2)}} \tag{77}$$

(this solution is unique). If we calculate the generalization error and the training error from (76) and (77), respectively, we recover the exact expressions

$$E_{\rm g} = \frac{1}{\pi}\arccos\left[\frac{R_0+\eta t\sqrt{\frac{2}{\pi}}}{\sqrt{Q_0+2\eta t R_0\sqrt{\frac{2}{\pi}}+\eta^2 t^2\left[\frac{1}{\alpha}+\frac{2}{\pi}\right]}}\right] \tag{78}$$

$$E_{\rm t} = \frac{1}{2} - \frac{1}{2}\int Dy\ \mathrm{erf}\left[\frac{|y|[R_0+\eta t\sqrt{\frac{2}{\pi}}]+\frac{\eta t}{\alpha}}{\sqrt{2[Q_0-R_0^2+\frac{\eta^2 t^2}{\alpha}]}}\right] \tag{79}$$

Comparison of (76,77) with (71,72) 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 $E_{\rm g}$) 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 $E_{\rm t}$, however, is less trivial, since here the disordered nature of the learning dynamics, leading to non-Gaussian distributions, is truly relevant.

4.2 On-Line Hebbian Learning

We next insert the Hebbian recipe $\mathcal{G}[x,y] = \mathrm{sgn}[y]$ into the on-line equations (38,39). Direct analytical solution of these equations, or a demonstration that they are solved by the exact result (69,70), although not ruled out, has not yet been achieved by us. The reason is that here one has conditionally Gaussian field distributions only in special limits. Numerical solution is in principle straightforward, but will be quite CPU intensive (see also a subsequent section). 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 (73) 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.

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

(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:

$$Q = Q_0 + 2\eta t R_0\sqrt{\frac{2}{\pi}} + \eta^2 t + \eta^2 t^2\left[\frac{1}{\alpha}+\frac{2}{\pi}\right] \qquad\qquad R = R_0 + \eta t\sqrt{\frac{2}{\pi}} \tag{80}$$

Comparing (80) with (69) 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

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

Multiplication of this equation by x followed by integration over x, together with usage of 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:

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

Comparison with (73) 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 $\hat{M}[k|y]=\int dx\ e^{-ikx}M[x|y]$:

$$\frac{d}{dt}\log\hat{P}[k|y]=\frac{1}{\alpha}\left[e^{-i\eta k\ \mathrm{sgn}(y)}-1\right]-i\eta k y\sqrt{\frac{2}{\pi}}-\frac{1}{2}\eta^2k^2-\frac{ik\eta^2 t}{\alpha\hat{P}[k|y]\sqrt{qQ-R^2}}\int Dz\ z\ \frac{\hat{M}[k+iBz|y]}{\hat{M}[iBz|y]}\qquad(81)$$

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

$$\hat{P}[k|y]=\int Dz\ \frac{\hat{M}[k+iBz|y]}{\hat{M}[iBz|y]}\qquad(82)$$

$$\frac{\eta^2t^2}{\alpha^2}+\int Dy\int dx\ P[x|y][x-\overline{x}(y)]^2+(1-\frac{1}{\alpha})(qQ-R^2)=\left[2Q(1-q)+\frac{1}{B^2}\right]\int DyDz\ \frac{\partial^2}{\partial z^2}\log\hat{M}[iBz|y]\qquad(83)$$

Since the fields x grow linearly in time (see our expression for $\overline{x}(y)$) the equations (81,83,82) 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 (81), followed by inserting $k\to k/t$ and by taking the limit $t\to\infty$, produces:

$$\log\hat{Q}_\infty[k|y]=-ikv(y)-\frac{i\eta^2k}{\alpha}\int_0^1du\ \lim_{t\to\infty}\frac{t}{\sqrt{qQ-R^2}}\int Dz\ z\ \frac{\hat{M}[uk/t+iBz|y]}{\hat{Q}_\infty[uk|y]\hat{M}[iBz|y]}\qquad(84)$$

with the functional saddle-point equation

$$\hat{Q}[k|y]=\int Dz\ \frac{\hat{M}[k/t+iBz|y]}{\hat{M}[iBz|y]}\qquad(85)$$

The rescaled asymptotic system (84,85) admits the solution

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

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

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

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

Inspection shows that these four asymptotic equations are solved by

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

so that

$$\lim_{t\to\infty}\hat{P}_t[k/t|y] = e^{-ik\eta\left[\alpha^{-1}\ \mathrm{sgn}(y)+y\sqrt{2/\pi}\right]-\frac{1}{2}\eta^2k^2/\alpha} \tag{86}$$

Comparison with (69,70) shows that this prediction (86) 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 conditionally-Gaussian form of the field distribution both at $t=0$ and at $t=\infty$, it would appear to make sense for us to approximate $P[x|y]$ and $M[x|y]$ by simple conditionally Gaussian distributions at any time:

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

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

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

$$\Delta^2 = (Q-R^2)[1+\mathcal{O}(t^{-1})] \qquad (t\to\infty)$$

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 (80), one obtains

$$E_{\rm g} = \frac{1}{\pi}\arccos\left[\frac{R}{\sqrt{Q}}\right] \qquad\qquad E_{\rm t} = \frac{1}{2}-\frac{1}{2}\int Dy\ \mathrm{erf}\left[\frac{|y|R+\eta t/\alpha}{\Delta\sqrt{2}}\right] \tag{87}$$

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

$$P(x) = \frac{e^{-\frac{1}{2}[x+\frac{\eta t}{\alpha}]^2/(\Delta^2+R^2)}}{2\sqrt{2\pi(\Delta^2+R^2)}}\left[1-\mathrm{erf}\left(\frac{R[x+\frac{\eta t}{\alpha}]}{\Delta\sqrt{2(\Delta^2+R^2)}}\right)\right]$$
$$+\frac{e^{-\frac{1}{2}[x-\frac{\eta t}{\alpha}]^2/(\Delta^2+R^2)}}{2\sqrt{2\pi(\Delta^2+R^2)}}\left[1+\mathrm{erf}\left(\frac{R[x-\frac{\eta t}{\alpha}]}{\Delta\sqrt{2(\Delta^2+R^2)}}\right)\right] \tag{88}$$

In figure 3 we compare the predictions for the generalization and training errors (87) 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 figure 4, 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 (73) 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 (88). The result is shown in figure 5, 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.

5 General Approximation Schemes

All three approximation schemes presented in this section aim at providing alternatives to calculating the effective measure $M[x|y]$ at each time step from the functional saddle-point equation. Since this calculation cannot (yet) be done analytically, it constitutes a significant numerical obstacle in working out the predictions of our theory. Each scheme preserves both normalisation and symmetries of the probability density $P[x, y]$ and its marginals, as well as the relation $\int dx\ P[x|y]\Phi[x, y] = 0$ for all y. In the first two approximation schemes, a large α expansion and a conditionally-Gaussian saddle-point approximation, all Gaussian integrals representing the disorder in the problem can be done analytically; this leads to a significant reduction in CPU time when solving our equations numerically (especially the large α approximation is extremely simple and fast, as it does not even involve a saddle-point equation for q). We only work out the equations for on-line learning; the batch laws follows as usual upon expanding the equations in powers of η and retaining only the linear terms.

5.1 Large α Approximation

Our first approximation scheme is obtained upon taking into account the finite nature of the training set (i.e. the disordered nature of the dynamics) in first non-trivial order. The amount of disorder is effectively measured by the parameter B, or, equivalently, by the deviation of the value of the spin-glass order parameter q from its naive value R^2/Q. Putting $B = 0$ in the saddle-point equation (45) immediately gives $\lim_{B\to 0} M[x|y] = P[x|y]$, so we write

$$M[x|y] = P[x|y]\ [1 + \sum_{\ell>0} B^\ell m_\ell[x|y]], \qquad \int dx\ P[x|y] m_\ell[x|y] = 0 \tag{89}$$

Upon inserting (89) as an *ansatz* into the saddle-point equation (45) , one easily shows that

$$M[x|y] = P[x|y]\ e^{-\frac{1}{2}B^2[x-\overline{x}(y)]^2 + \frac{1}{2}B^2\left[\overline{x^2}(y) - \overline{x}(y)^2\right] + \mathcal{O}(B^3)} \tag{90}$$

with the abbreviations

$$\overline{x}(y) = \int dx\ P[x|y] x \qquad \overline{x^2}(y) = \int dx\ P[x|y] x^2$$

(the second $\mathcal{O}(B^2)$ term in the exponent of (90), being independent of x, just reflects the normalisation requirements). This result enables us, in turn, to expand the function $\Phi[x, y]$ which controls the non-trivial term in our diffusion equation for $P[x|y]$. Note that from the definition of B it follows that $Q(1-q) = \frac{1}{2}B^{-2}[\sqrt{1+4B^2(Q-R^2)}-1]$, which gives

$$\Phi[x, y] = \frac{x - \overline{x}(y)}{Q - R^2} + \mathcal{O}(B^2)$$

With this expression we can write our approximate equations in explicitly closed form (i.e. without any remaining saddle-point equations). The relevant scalar functions become

$$U = \frac{\langle \mathcal{G}[x, y][x - \overline{x}(y)]\rangle}{Q - R^2} \qquad V = \langle x\mathcal{G}[x, y]\rangle \qquad W = \langle y\mathcal{G}[x, y]\rangle \qquad Z = \langle \mathcal{G}^2[x, y]\rangle \tag{91}$$

For on-line learning we find:

$$\frac{d}{dt}Q = 2\eta V + \eta^2 Z \qquad \frac{d}{dt}R = \eta W \tag{92}$$

$$\frac{d}{dt}P[x|y] = \frac{1}{\alpha}\int dx' P[x'|y]\,[\delta[x-x'-\eta\mathcal{G}[x',y]]-\delta[x-x']] - \eta\frac{\partial}{\partial x}\left\{P[x|y]\,[U(x-Ry)+Wy]\right\}$$
$$+\frac{1}{2}\eta^2 Z\frac{\partial^2}{\partial x^2}P[x|y] - \eta\left[\frac{V-RW}{Q-R^2}-U\right]\frac{\partial}{\partial x}\left\{P[x|y][x-\overline{x}(y)]\right\} \tag{93}$$

From the solution of the above equations follow, as always, the training- and generalization errors $E_{\rm t} = \int Dydx\ P[x|y]\theta[-xy]$ and $E_{\rm g} = \pi^{-1}\arccos[R/\sqrt{Q}]$. The resulting theory is obviously exact in the limit $\alpha\to\infty$, by construction.

5.2 Conditionally-Gaussian Approximation

Our basic idea here is a variational approach to solving the functional saddle-point problem (valid for any α), i.e. to carry out the functional extremisation only within the restricted family of conditionally Gaussian measures $M[x|y]$ (which, together with q, characterises the saddle-point):

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

Note that this does not imply the stronger statement that $P[x|y]$ itself is taken to be of a conditionally-Gaussian form (as in the case of the approximation used for on-line Hebbian learning). Extremisation of the original replica-symmetric functional $\Psi[q,\{M\}]$ within the conditionally-Gaussian family of functions results in the requirement that the two y-dependent moments $\overline{x}(y)$ and $\sigma^2(y)$ be given by

$$\overline{x}(y) = \int dx\ xP[x|y], \qquad \Delta^2(y) = \int dx\ x^2P[x|y] - \overline{x}^2(y) = \sigma^2(y)+B^2\sigma^4(y)$$

Now we can again calculate all relevant averages which involve the effective measure $M[x|y]$ exactly. In particular:

$$\langle x\rangle_\star = \overline{x}(y)+zB\sigma^2(y) \qquad B=\frac{\sqrt{qQ-R^2}}{Q(1-q)} \qquad \Phi[x,y] = \frac{e^{-\frac{1}{2}[x-\overline{x}(y)]^2/\Delta^2(y)}}{\Delta(y)\sqrt{2\pi}P[x|y]}\frac{(x-\overline{x}(y))\sigma^2(y)}{Q(1-q)\Delta^2(y)}$$

For on-line learning this results in the following approximated theory:

$$U = \int DyDu\left\{\frac{u\sigma^2(y)\mathcal{G}[\overline{x}(y)+u\Delta(y),y]}{Q(1-q)\Delta(y)}\right\}$$

$$V=\langle x\mathcal{G}[x,y]\rangle \qquad W=\langle y\mathcal{G}[x,y]\rangle \qquad Z=\langle\mathcal{G}^2[x,y]\rangle \tag{94}$$

$$\frac{d}{dt}Q = 2\eta V+\eta^2 Z \qquad \frac{d}{dt}R=\eta W \tag{95}$$

$$\frac{d}{dt}P[x|y] = \frac{1}{\alpha}\int dx' P[x'|y]\,[\delta[x-x'-\eta\mathcal{G}[x',y]]-\delta[x-x']] - \eta\frac{\partial}{\partial x}\left\{P[x|y]\,[U(x-Ry)+Wy]\right\}$$
$$+\frac{1}{2}\eta^2 Z\frac{\partial^2}{\partial x^2}P[x|y] - \frac{\eta\sigma^2(y)\,[V-RW-(Q-R^2)U]}{\sqrt{2\pi}Q(1-q)\Delta^5(y)}\left[\Delta^2(y)-(x-\overline{x}(y))^2\right]e^{-\frac{1}{2}[x-\overline{x}(y)]^2/\Delta^2(y)} \tag{96}$$

The remaining order parameter q is calculated at each time-step by solving

$$\langle(x-Ry)^2\rangle + (qQ-R^2)(1-\frac{1}{\alpha}) = \left[2\frac{qQ-R^2}{Q(1-q)}+1\right]\int Dy\ \sigma^2(y)$$

From the solution of these equations follow the training- and generalization errors $E_{\rm t} = \int Dydx\ P[x|y]\theta[-xy]$ and $E_{\rm g} = \pi^{-1}\arccos[R/\sqrt{Q}]$.

5.3 Partially Annealed Approximation

In order to construct our third and final approximation we return to an earlier stage of the derivation of the present formalism, and rewrite the functional saddle-point equation in a form where the replica limit $n \to 0$ has not yet been taken, i.e.

$$\text{for all } x,y: \qquad P[x|y] = \frac{\int Dz\ M_n[x|y] e^{Bz[x-\overline{x}(y)]} \left[\int dx'\ M_n[x'|y] e^{Bz[x'-\overline{x}(y)]}\right]^{n-1}}{\int Dz \left[\int dx'\ M_n[x'|y] e^{Bz[x'-\overline{x}(y)]}\right]^n}$$

with $\overline{x}(y) = \int dx\ xP[x|y]$. In our full (quenched disorder) calculation we find ourselves with the effective measure $M[x|y] = \lim_{n\to 0} M_n[x|y]$. In contrast, an alternative calculation, whereby the quenched average over all training sets would have been replaced by an annealed average over all training sets, would have led us to the value $n = 1$ rather than $n = 0$: $M[x|y] = M_1[x|y]$. We can now define in a natural way an annealed approximation of our theory upon replacing the complicated $n = 0$ functional saddle-point equation (45) by the much simpler $n = 1$ version:

$$P[x|y] = \frac{\int Dz\ M[x|y] e^{Bz[x-\overline{x}(y)]}}{\int Dz \int dx'\ M[x'|y] e^{Bz[x'-\overline{x}(y)]}}$$

The z-integrations can immediately be carried out, and the resulting equation solved for $M[x|y]$, giving:

$$M[x|y] = \frac{P[x|y]\ e^{-\frac{1}{2}B^2[x-\overline{x}(y)]^2}}{\int dx'\ P[x'|y]\ e^{-\frac{1}{2}B^2[x'-\overline{x}(y)]^2}}, \tag{97}$$

Averages involving the effective measure $M[x|y]$ are thus written explicitly in terms of $P[x|y]$, and we are left with the following approximate theory:

$$U = \langle \Phi[x,y]\mathcal{G}[x,y]\rangle \qquad V = \langle x\mathcal{G}[x,y]\rangle \qquad W = \langle y\mathcal{G}[x,y]\rangle \qquad Z = \langle \mathcal{G}^2[x,y]\rangle \tag{98}$$

$$\frac{d}{dt}Q = 2\eta V + \eta^2 Z \qquad \frac{d}{dt}R = \eta W \tag{99}$$

$$\frac{d}{dt}P[x|y] = \frac{1}{\alpha}\int dx' P[x'|y]\,[\delta[x-x'-\eta\mathcal{G}[x',y]]-\delta[x-x']] - \eta\frac{\partial}{\partial x}\left\{P[x|y]\,[U(x-Ry)+Wy]\right\}$$
$$+\frac{1}{2}\eta^2 Z\frac{\partial^2}{\partial x^2}P[x|y] - \eta\left[V-RW-(Q-R^2)U\right]\frac{\partial}{\partial x}\left\{P[x|y]\Phi[x,y]\right\} \tag{100}$$

with

$$\Phi[X,y] = \frac{1}{Q(1-q)}\int Dz \left\{\frac{\int dx\ P[x|y]\ e^{-\frac{1}{2}[B(x-\overline{x}(y))-z]^2 - \frac{1}{2}[B(X-\overline{x}(y))-z]^2}(X-x)}{\left[\int dx\ P[x|y]\ e^{-\frac{1}{2}[B(x-\overline{x}(y))-z]^2}\right]^2}\right\}$$

As always, $B = \sqrt{qQ-R^2}/Q(1-q)$. The remaining spin-glass order parameter q is calculated at each time-step by solving

$$\langle (x-Ry)^2\rangle + (qQ-R^2)(1-\frac{1}{\alpha}) = \left[2(qQ-R^2)^{\frac{1}{2}} + \frac{1}{B}\right]\int DyDz\ z\left\{\frac{\int dx\ P[x|y]\ e^{-\frac{1}{2}[B(x-\overline{x}(y))-z]^2}x}{\int dx\ P[x|y]\ e^{-\frac{1}{2}[B(x-\overline{x}(y))-z]^2}}\right\}$$

From the solution of the above equations follow the training- and generalization errors $E_{\rm t} = \langle\theta[-xy]\rangle$ and $E_{\rm g} = \pi^{-1}\arccos[R/\sqrt{Q}]$. It should be emphasised that the present approximation is not equivalent to (and should be more accurate than) a full annealed treatment of the disorder in the problem; the latter would have affected not only the equation for $M[x|y]$ but also the saddle-point equation for q (hence the name *partially* annealed approximation).

6 Non-Hebbian Rules: Theory versus Simulations

Henceforth we will always assume initial states with specified values for R_0 and Q_0 but without correlations with the training set, i.e.

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

This implies that the student could initially have some knowledge of the rule to be learned, if we wish, but will never know beforehand about the composition of the training set. We will inspect the learning dynamics generated upon using two of the most common non-Hebbian (error-correcting) learning rules:

$$\begin{array}{ll} \text{Perceptron}: & \mathcal{G}[x,y] = \ \text{sgn}(y)\theta[-xy] \\ \text{AdaTron}: & \mathcal{G}[x,y] = |x|\ \text{sgn}(y)\theta[-xy] \end{array} \tag{101}$$

Note that in the case of AdaTron learning the cases $\eta \leq 1$ and $\eta > 1$ give rise to qualitatively different behaviour of the first term in the diffusion equation (39). For $\eta < 1$ the learning process, aiming at the situation where $xy > 0$ never occurs, remedies inappropriate student fields by slowly moving them towards (but not immediately across) the decision boundary. For $\eta > 1$ the adjustments made to the student fields could move them well into the region at the other side of the decision boundary. The case $\eta = 1$ is special, in that changes to the student fields tend to move them precisely onto the decision boundary. The student field distribution consequently develops a δ-peak at the origin, in perfect agreement with what can be observed in numerical simulations (see e.g. the graphs referring to on-line AdaTron learning with $\eta = 1$ in figures 1 and 2):

$$\eta = 1: \qquad \frac{d}{dt}P[x|y] = \frac{1}{\alpha}\left\{\delta(x)\int dx'\ \theta[-x'y]P[x'|y] - P[x|y]\theta[-xy]\right\} + \ldots$$

In fact the same occurs for all $\eta \leq 1$: about half of the probability weight of $P[x|y]$ will in due course become concentrated in an increasingly thin ridge along the decision boundary $x = 0$. This is illustrated in figure 6, for $\eta = \frac{1}{2}$. Since such a singular behaviour (although in principle accurately described by our equations) will be difficult to reproduce when solving the equations numerically, using finite spatial resolution, we will in this paper only deal with the case of $\eta > 1$ for AdaTron learning.

6.1 Large α and Conditionally-Gaussian Approximations

Our first approximated theory (the large α approximation) is very simple, with neither saddle-point equations to be solved nor nested integrations. As a result, numerical solution of the macroscopic equations is straightforward and fast. In figures 7 (on-line perceptron learning) and 8 (on-line Adatron learning) we compare the results of solving the coupled equations (91,92,93) numerically for finite values of α, plotting the generalisation- and training errors as functions of time, with results obtained from performing numerical simulations. As could have been expected, the large α approximation under-estimates the amount of disorder in the learning process, which immediately translates into under-estimation of the gap between $E_{\rm t}$ and $E_{\rm g}$ (which is its fingerprint). It is also clear from these figures that, although at any given time the quality of the predictions of this approximation does improve when α increases (as indeed it should), and although there is surely qualitative agreement, reliably accurate quantitative statements on the values of the training- and generalisation errors are confined to the regime $\eta t \leq \alpha$. Yet, surprisingly, the agreement obtained is very good, even for $\eta t > \alpha$. Apparently the present approximation does still capture the main characteristics of the (non-Gaussian) joint field distribution. This is illustrated quite clearly and explicitly in figures 9 and 10, where we compare for a fixed time $t = 10$ the student and teacher fields as measured during numerical simulations (for $N = 10,000$, drawn as dots in the (x, y) plane) for the $p = \alpha N$ questions $\boldsymbol{\xi}^\mu$ in the training set $\tilde{D}$, to the theoretical predictions for the joint field distribution $P[x, y]$ (drawn as contour plots). We will not at this stage attempt to explain the surprising effectiveness of the large α approximation for small values of α (note that figures 7 and 8 even suggest an increase in accurateness

as α is lowered below $\alpha = 1$). This would require a systematic mathematical analysis of the non-linear diffusion equation (93), which we consider to be beyond the scope of the present paper.

The conditionally-Gaussian approximation again involves no nested integrals, and its equations can therefore still be solved numerically in a reasonably fast way, but it does already require the solution (at each infinitesimal time step) of a scalar saddle-point equation to determine the spin-glass order parameter q. Approximations of this type work extremely well for the simple Hebbian learning rules, as we have seen earlier. However, numerical solution of the coupled equations (94,95,96) shows quite clearly that for the more sophisticated non-Hebbian rules such as Perceptron and AdaTron, which are of an error correcting nature (i.e. where where changes are made only when student and teacher disagree), the conditionally-Gaussian approximation is less accurate than the previously investigated large α approximation, in spite of the fact that the latter involved much simpler equations. Apparently the generally non-Gaussian nature of the conditional distribution $P[x|y]$, and thereby of the measure $M[x|y]$, is of crucial importance. It is not good enough to try getting away with allowing the y-dependent averages $\overline{x}(y)$ and variances $\Delta(y)$ to be non-trivial functions. With conditionally-Gaussian measures $M[x|y]$ it turns out that generating the right width of the conditional distributions $P[x|y]$ inevitably introduces tails for $P[x|y]$ which spill into the $xy < 0$ region, which are found to be absent in error-correcting learning rules such as Perceptron and Adatron. This picture is consistent with figures 9 and 10, where we can observe that for any fixed value of the teacher field y the remaining marginal distribution for x is generally not symmetric around its (y-dependent) average. We conclude that the conditionally-Gaussian approximation is generally inferior to the large α approximation. We will not waste paper by producing large numbers of graphs to illustrate this explicitly and comprehensively, but we will rather draw the conditionally-Gaussian predictions together with those of the other approximations and of the full theory, by way of illustration.

6.2 Partially Annealed Approximation and Full Equations

The partially annealed approximation and the full theory are both expected to improve upon the large α approximation (note that the partially annealed approximation can be seen as an improved version of the large α approximation, similar in structure but valid also for small α, i.e. large B). Although the partially annealed approximation does not involve a functional saddle-point equation to be solved (which improves numerical speed), it shares with the full theory the appearance of nested (Gaussian) integrals, namely those appearing in the function $\Phi[x, y]$ and in the saddle-point equation for q. Thus, solution of both the full theory and of the partially annealed approximation involves a significant amount of CPU time (avoiding standard instabilities of discretised diffusion equations sets further limits on the maximum size of the time discretisation, dependent on the field resolution [17]), which implies that we have to reduce our ambition and restrict the number of experiments to a few typical ones.

We will thus investigate two examples, both with $\alpha = 1$: on-line Perceptron learning with $\eta = \frac{1}{2}$, and on-line AdaTron learning with $\eta = \frac{3}{2}$. We solve numerically the full equations of our theory, i.e. the macroscopic dynamical laws (38,39) with the order parameters calculated at each time step by solving (44,45), and show in figure 11 the training and generalisation errors as functions of time together with the corresponding values as measured during numerical simulations, with systems of size $N = 10,000$. In addition, we plot in the same picture, for comparison, the training- and generalisation errors obtained by numerical solution of the three approximated theories as derived in the previous section. In comparing curves we have to take into account that those describing the large α approximation were generated upon solving the diffusion equation with a significantly higher numerical field resolution ($\Delta x = 0.015$) than the others (where we used $\Delta x = 0.05$), because of CPU limitations. A restricted field resolution is likely to be more critical at large times, where the probability weight in the $xy < 0$ region, responsible for the residual error and for the non-stationarity of the dynamics, is highly concentrated close to the decision boundary $x = 0$. Especially for large times, we should therefore expect the full theory, the conditionally-Gaussian approximation, and the partially annealed approximation to all three perform better in reality than what is suggested by the numerical solutions of their equations as shown in figure 11. This is particularly true for AdaTron learning, where even for $\eta > 1$ (where we do not expect to observe a δ-singularity) the field distributions still tend to develop a jump discontinuity at $x = 0$. It turns out that the curves of the full theory and those of the partially annealed approximation are

very close (virtually on top of one another for the case of Perceptron learning) in figure 11; apparently for the learning times considered here there is no real need to evaluate the full theory.

Finally, we show in figure 12 for both the full theory and for the simulation experiments the two distributions $P^{\pm}(x) = \int dy\ P[x,y]\theta[\pm y]$ for the student fields, given a specified sign of the teacher field y (and thus a given teacher output), corresponding to the same experiments. Note that $P(x) = P^{+}(x) + P^{-}(x)$. The pictures in figure 12 again illustrate quite clearly the difference between learning with restricted training sets and learning with infinite training sets: in the former case the desired agreement $xy > 0$ between student and teacher is achieved by a qualitative *deformation* of $P[x|y]$, away from the initial Gaussian shape, rather than by adaptation of the first and second order moments.

Our restricted resolution numerics obviously have difficulty in reproducing the discontinuous behaviour of $P^{\pm}(x)$ near $x = 0$ for on-line Adatron learning (as expected), which explains why in this regime the simplest large α approximation (which can be numerically evaluated with almost arbitrarily high field resolution) appears to outperform the more sophisticated versions of the theory (which CPU limitations force us to evaluate with rather limited field resolution), according to figure 11.

We conclude from the results in this section that our full theory indeed gives an adequate description of the macroscopic process, and that the partially annealed approximation is almost equivalent in performance to the full theory. As mentioned before, the conditionally-Gaussian approximation performs generally poorly (except, as we have seen earlier, for the simple Hebian rule). Which of the remaining three versions of our theory to use in practice will clearly depend on the accuracy constraints and available CPU time of the user, with the full theory at the higher end of the market (in principle very accurate, but almost too CPU expensive to work out and exploit properly), with the large α approximation on the lower end (reasonably acurate, but very cheap), and with the annealed approximation as a sensible compromise in between these two.

7 Discussion

In this paper we have shown how the formalism of dynamical replica theory (see e.g. [13]) can be successfully employed to construct a general theory which enables one to predict the evolution of the relevant macroscopic performance measures for supervised (on-line and batch) learning in layered neural networks, with randomly chosen but restricted training sets, i.e. for finite $\alpha = p/N$ where weight updates are carried out by sampling with repetition. In this case the student nodes local fields are no longer described by (multivariate) Gaussian distributions and the 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 the student and teacher local fields is itself taken to be a dynamical order parameter, in addition to the conventional observables Q and R representing overlaps between the student-student and student-teacher vectors respectively. The new order parameter set $\{Q,R,P\}$, in turn, enables one to monitor the generalization error $E_{\rm g}$ as well as 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 generally 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 (in the $N \to \infty$ limit), providing 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 can be interpreted as employing a maximum entropy 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, allowed us to perform a critical test of the theory. For batch Hebbian learning we demonstrate explicitly that our theory is fully exact. For on-line Hebbian

[4]The reason why the replica formalism is 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.

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 moments of the conditional field distributions $P[x|y]$ (for any $y \in \Re$), 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 essentially perfectly.

No exact benchmark solution is available for non-Hebbian (i.e. non-trivial) learning rules, leaving numerical simulations as the only yardstick against which to test our theory. Motivated by the need to solve a functional saddle-point equation at each time step in the full theory, and by the presence of nested integrations, we have constructed a number of systematic approximations to the original equations. We have compared the predictions of the full theory and of the three approximation schemes with one another and with the results obtained upon performing numerical simulations of non-linear learning rules, such as Perceptron and AdaTron, in large perceptrons (of size $N = 10,000$), with various values of learning rates η and relative training set sizes α. One of the approximations, a conditionally-Gaussian saddle-point approximation in the spirit of the particular approximation that was found to work perfectly for Hebbian learning, turned out to perform badly for general non-Hebbian rules. The other two approximations, the large α approximation and the partially annealed approximation, each have their specific usefulness; the former is extremely simple and fast, whereas the latter is overall more accurate, but more expensive in its CPU requirements (so that in practice its true accurateness cannot always be realised). Yet, the large α approximation still works remarkably well, even for small α, in spite of it being so simple that it can be written as a fully explicit set of equations for Q, R and the joint field distribution $P[x,y]$ only. The observed accuracy of these simple equations in the small α regime suggests that for $\alpha \to 0$ the leading term in the diffusion equation for $P[x|y]$ is the first term in the right-hand side, which reflects the direct effect of pattern recycling, and which indeed has not been approximated.

We believe that our theory offers an efficient tool with which to analyse and predict the outcome of learning processes in single-layer networks. In particular, for those who are primarily interested in the progress and the outcome of learning processes there is no real need to understand the full details of the derivation; one can simply adopt the macroscopic laws (or one of the two appropriate approximations, to save CPU time) as a starting point, and just apply them to the learning rules as hand. In the applications worked out in this paper (Hebbian learning, Perceptron learning and AdaTron learning) our formalism has been found to be either exact or an excellent approximation. It is not realistic to expect that simpler theories can be found with a similar level of accuracy. While putting the finishing touch to this manuscript a preprint was communicated [18] in which the authors apply the cavity method to the present problem. They manage to keep their theory relatively simple by restricting themselves to batch learning and to gradient descent learning rules, and by applying their theory only to a linear learning rule. Here also the present theory would have been both simpler and exact. A fully exact theory for both on-line and batch learning and for arbitrary learning rules can be constructed [19] using a suitable adaptation of the generating functional methods as in [10], but in describing transients it will be more complicated than the present one, as it will be built around macroscopic observables with two time arguments (correlation- and response functions) and will take the form of an effective single weight process with coloured stochastic noise and retarded self-interactions. It will, however, be interesting to see the connection between the generating functional theory and the present dynamical replica formalism.

The present study opens up new possibilities for considering unrealizable learning scenarios, either due to structural limitations or due to noise, which require some sort of regularization. The examination of regularization techniques in such scenarios, which is of great practical significance, was out of reach so far as they come into effect only where the error-surface is fixed by having a fixed example set. It turns out that the case of noisy teachers can be studied with an appropriate extension of the present formalism [21], involving a joint distribution of three rather than two fields (namely those of student, 'clean' teacher, and 'noisy' teacher). Generalization to multi-layer networks (with a finite number of hidden nodes) is also straightforward [20], although numerically intensive. At a more fundamental level one could explore the effects 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). Last, but not least, our theory would simplify

significantly if one could find a more explicit solution of the functional saddle-point equation (131), enabling us to express the function $\Phi[x,y]$ directly in terms of our order parameters.

Acknowledgements

It is our pleasure to thank Yuan-sheng Xiong and Charles Mace for valuable discussions. We also acknowledge support by EPSRC (Grant GR/L52093), the British Council (British-German Academic Research Collaboration Programme project 1037), and the London Mathematical Society (grant 4415).

References

[1] Kinzel W and Rujan P 1990 *Europhys. Lett.* **13** 473

[2] Kinouchi O and Caticha N 1992 *J. Phys. A: Math. Gen.* **25** 6243

[3] Biehl M and Schwarze H 1992 *Europhys. Lett.* **20** 733

[4] Biehl M and Schwarze H 1995 *J. Phys. A: Math. gen.* **28** 643

[5] Saad D and Solla S 1995 *Phys. Rev.* **E52** 4225; *Phys. Rev.* **L74** 4337

[6] Mace CWH and Coolen ACC 1998 *Statistics and Computing* **8** 55

[7] Saad D (Ed) 1998 *On-line Learning in Neural Networks* (Cambridge: Cambridge University Press)

[8] Barber D, Saad D and Sollich P 1996 *Europhys. Lett.* **34** 151

[9] Sollich P and Barber D 1998 in *Advances in Neural Information Processing Systems* Jordan MI, Kearns MJ and Solla SA (Eds) (MIT Press, Cambridge, MA) Vol. 10, p 385

[10] Horner H 1992 *Z. Phys.* **B86** 291; *Z. Phys.* **B87** 371

[11] Krogh A and Hertz JA 1992 *J. Phys. A: Math. Gen.* **25** 1135

[12] Rae HC, Sollich P and Coolen ACC 1999 *J. Phys. A: Math. Gen.* **32** 3321

[13] Coolen ACC, Laughton SN and Sherrington D 1996 *Phys. Rev.* **B53** 8184

[14] Coolen ACC, Saad D and Xiong YS 2000 *submitted to Europhys. Lett.*

[15] Mace CWH and Coolen ACC 2000 in em Advanced in Neural Information Processing Systems Solla SA, Leen TK and Müller K (Eds) (MIT Press, Cambridge, MA) Vol. 12 (in press)

[16] Mézard M, Parisi G and Virasoro MA 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)

[17] Press WH, Flannery BP, Teukolsky SA and Vetterling WT 1988 *Numerical Recipes in C* (Cambridge: U.P.)

[18] Wong KY, Li S and Tong YW 1999 *preprint cond-mat/9909004*

[19] Heimel JA and Coolen ACC 2000 *in preparation*

[20] Coolen ACC, Saad D and Xiong YS 2000 *submitted to Europhys. Lett.*

[21] Mace CWH and Coolen ACC 2000 in em Advanced in Neural Information Processing Systems Solla SA, Leen TK and Müller K (Eds) (MIT Press, Cambridge, MA) Vol. 12 (in press)

A Diffusion Matrix Elements of the Macroscopic Fokker-Planck Equation

Here we show that for the observables (20) the diffusion matrix elements $G^{\star\star\star}_{\mu\nu}$ in the Fokker-Planck equation (19) vanish for $N\to\infty$. Our observables will consequently obey deterministic dynamical laws. Calculating diffusion terms associated with $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$ is trivial:

$$\left[\begin{array}{c} G^{\rm onl}_{QQ}[\ldots] \\ G^{\rm onl}_{QR}[\ldots] \\ G^{\rm onl}_{RR}[\ldots] \end{array}\right] = \lim_{N\to\infty}\frac{\eta^2}{N}\int dxdy\ P[x,y]\ \mathcal{G}^2[x,y]\left[\begin{array}{c} 4x^2 \\ 2xy \\ y^2 \end{array}\right] = 0$$

$$\left[\begin{array}{c} G^{\rm bat}_{QQ}[\ldots] \\ G^{\rm bat}_{QR}[\ldots] \\ G^{\rm bat}_{RR}[\ldots] \end{array}\right] = \lim_{N\to\infty}\frac{\eta^2}{N}\left[\begin{array}{c} 4\left\{\int dxdy\ P[x,y]\ x\mathcal{G}[x,y]\right\}^2 \\ 2\left\{\int dxdy\ P[x;y]\ x\mathcal{G}[x,y]\right\}\left\{\int dxdy\ P[x;y]\ y\mathcal{G}[x,y]\right\} \\ \left\{\int dxdy\ P[x;y]\ y\mathcal{G}[x,y]\right\}^2 \end{array}\right] = 0$$

We next turn to diffusion terms with one occurrence of $P[x,y;\boldsymbol{J}]$. Here we repeatedly build on the cornerstone assumption that all fields $\boldsymbol{J}\cdot\boldsymbol{\xi}$ and $\boldsymbol{B}\cdot\boldsymbol{\xi}$ are of order unity (which is clear from numerical simulations, and will be supported self-consistently by the equations resulting from our theory), in combination with two simple scaling consequences of the random composition of $\tilde{D}$, as $N\to\infty$:

$$\boldsymbol{\xi}\in\tilde{D}:\quad \frac{1}{p}\sum_{\boldsymbol{\xi}'\in\tilde{D}}\delta_{\boldsymbol{\xi}\boldsymbol{\xi}'} = p^{-1}+\mathcal{O}(p^{-2}) \qquad\qquad \frac{1}{p^2}\sum_{\boldsymbol{\xi}\in\tilde{D}}\sum_{\boldsymbol{\xi}'\in\tilde{D}}[1-\delta_{\boldsymbol{\xi}\boldsymbol{\xi}'}]\ |\boldsymbol{\xi}\cdot\boldsymbol{\xi}'| = \mathcal{O}(N^{\frac{1}{2}}) \tag{102}$$

For on-line learning we find:

$$\left[\begin{array}{c} G^{\rm onl}_{Q,P[x,y]}[\ldots] \\ G^{\rm onl}_{R,P[x,y]}[\ldots] \end{array}\right] = -\lim_{N\to\infty}\frac{\eta^2}{N}\frac{\partial}{\partial x}\left\langle \langle\!\langle\ \mathcal{G}^2[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\left[\begin{array}{c}2\boldsymbol{J}\cdot\boldsymbol{\xi}\\ \boldsymbol{B}\cdot\boldsymbol{\xi}\end{array}\right](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}']\ \rangle\!\rangle_{\tilde{D}}\rangle_{\tilde{D}}\right\rangle_{\rm QRP;t}$$

$$= -\eta^2\frac{\partial}{\partial x}\lim_{N\to\infty}\left\langle \frac{1}{N}\langle\!\langle\ [1-\delta_{\boldsymbol{\xi}\boldsymbol{\xi}'}]\mathcal{G}^2[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\left[\begin{array}{c}2\boldsymbol{J}\cdot\boldsymbol{\xi}\\ \boldsymbol{B}\cdot\boldsymbol{\xi}\end{array}\right](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\ \rangle\!\rangle_{\tilde{D}}\rangle_{\tilde{D}}\right.$$

$$\left. + \mathcal{G}^2[x,y]\left[\begin{array}{c}2x\\ y\end{array}\right]\langle\!\langle\ \delta_{\boldsymbol{\xi}\boldsymbol{\xi}'}\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\ \rangle\!\rangle_{\tilde{D}}\rangle_{\tilde{D}}\ \right\rangle_{\rm QRP;t}$$

$$= -\eta^2\frac{\partial}{\partial x}\lim_{N\to\infty}\left\langle \mathcal{O}(N^{-\frac{1}{2}})+\mathcal{O}(N^{-1})\right\rangle_{\rm QRP;t} = 0$$

For batch learning we find:

$$\left[\begin{array}{c} G^{\rm bat}_{Q,P[x,y]}[\ldots] \\ G^{\rm bat}_{R,P[x,y]}[\ldots] \end{array}\right] = -\lim_{N\to\infty}\frac{\eta^2}{N}\frac{\partial}{\partial x}\int dx'dy'P[x',y']\mathcal{G}[x',y']\left[\begin{array}{c}2x'\\ y'\end{array}\right]$$

$$\times\left\langle \langle\!\langle\ \mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}']\ \rangle\!\rangle_{\tilde{D}}\rangle_{\tilde{D}}\right\rangle_{\rm QRP;t}$$

$$= -\eta^2\frac{\partial}{\partial x}\int dx'dy'P[x',y']\mathcal{G}[x',y']\left[\begin{array}{c}2x'\\ y'\end{array}\right]\lim_{N\to\infty}\left\langle \mathcal{G}[x,y]\langle\!\langle\ \delta_{\boldsymbol{\xi}\boldsymbol{\xi}'}\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\ \rangle\!\rangle_{\tilde{D}}\rangle_{\tilde{D}}\right.$$

$$\left. +\frac{1}{N}\langle\!\langle\ [1-\delta_{\boldsymbol{\xi}\boldsymbol{\xi}'}]\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}']\ \rangle\!\rangle_{\tilde{D}}\rangle_{\tilde{D}}\ \right\rangle_{\rm QRP;t}$$

$$= -\eta^2 \frac{\partial}{\partial x} \lim_{N\to\infty} \left\langle \mathcal{O}(N^{-1}) + \mathcal{O}(N^{-\frac{1}{2}}) \right\rangle_{\rm QRP;t} = 0$$

The difficult terms are those where two derivatives of the order parameter function $P[x,y;\boldsymbol{J}]$ come into play. Here we have to deal separately with four distinct contributions, defined according to which of the vectors from the trio $\{\boldsymbol{\xi},\boldsymbol{\xi}',\boldsymbol{\xi}''\}$ are identical. For on-line learning we find:

$$G^{\rm onl}_{P[x,y],P[x',y']}[\ldots] = \lim_{N\to\infty} \frac{\eta^2}{N} \frac{\partial^2}{\partial x \partial x'}$$

$$\left\langle \langle\langle\langle\, \mathcal{G}^2[\boldsymbol{J}\cdot\boldsymbol{\xi}'',\boldsymbol{B}\cdot\boldsymbol{\xi}''](\boldsymbol{\xi}\cdot\boldsymbol{\xi}'')(\boldsymbol{\xi}'\cdot\boldsymbol{\xi}'')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\,\rangle_{\bar D}\rangle_{\bar D}\rangle_{\bar D} \right\rangle_{\rm QRP;t}$$

$$= \eta^2 \frac{\partial^2}{\partial x\partial x'} \lim_{N\to\infty} \Big\langle \;\; N\mathcal{G}^2[x,y]\delta[x'-x]\delta[y'-y]\;\langle\langle\langle\, \delta_{\boldsymbol{\xi}\boldsymbol{\xi}''}\delta_{\boldsymbol{\xi}'\boldsymbol{\xi}''}\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\,\rangle_{\bar D}\rangle_{\bar D}\rangle_{\bar D}$$

$$+\;\mathcal{G}^2[x',y']\;\langle\langle\langle\, \overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}''}\delta_{\boldsymbol{\xi}'\boldsymbol{\xi}''}(\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\,\rangle_{\bar D}\rangle_{\bar D}\rangle_{\bar D}$$

$$+\;\mathcal{G}^2[x,y]\;\langle\langle\langle\, \delta_{\boldsymbol{\xi}\boldsymbol{\xi}''}\overline{\delta}_{\boldsymbol{\xi}'\boldsymbol{\xi}''}(\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\,\rangle_{\bar D}\rangle_{\bar D}\rangle_{\bar D}$$

$$+\langle\langle\langle\, \overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}''}\overline{\delta}_{\boldsymbol{\xi}'\boldsymbol{\xi}''}\mathcal{G}^2[\boldsymbol{J}\cdot\boldsymbol{\xi}'',\boldsymbol{B}\cdot\boldsymbol{\xi}'']\frac{(\boldsymbol{\xi}\cdot\boldsymbol{\xi}'')(\boldsymbol{\xi}'\cdot\boldsymbol{\xi}'')}{N}\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\,\rangle_{\bar D}\rangle_{\bar D}\rangle_{\bar D} \Big\rangle_{\rm QRP;t}$$

$$= \eta^2 \frac{\partial^2}{\partial x\partial x'} \left\{ \lim_{N\to\infty} \left\langle \mathcal{O}(N^{-1}) + \mathcal{O}(N^{-\frac{1}{2}}) + \mathcal{O}(N^{-\frac{1}{2}}) \right\rangle_{\rm QRP;t} + \int dx''dy''\mathcal{G}^2[x'',y'']\;\mathcal{C}[x,y;x',y';x'',y''] \right\}$$

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

Similarly:

$$G^{\rm bat}_{P[x,y],P[x',y']}[\ldots] = \lim_{N\to\infty} \frac{\eta^2}{N} \frac{\partial^2}{\partial x \partial x'}$$

$$\left\langle \langle\langle \mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi}',\boldsymbol{B}\cdot\boldsymbol{\xi}'](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar D}\rangle_{\bar D}\;\langle\langle \mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}](\boldsymbol{\xi}\cdot\boldsymbol{\xi}')\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\rangle_{\bar D}\rangle_{\bar D} \right\rangle_{\rm QRP;t}$$

$$= \eta^2 \frac{\partial^2}{\partial x\partial x'} \lim_{N\to\infty} \Big\langle \left\{ \mathcal{G}[x,y]\langle\langle \delta_{\boldsymbol{\xi}\boldsymbol{\xi}'}\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar D}\rangle_{\bar D} + \langle\langle \overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi}',\boldsymbol{B}\cdot\boldsymbol{\xi}']\frac{\boldsymbol{\xi}\cdot\boldsymbol{\xi}'}{N}\delta[x-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar D}\rangle_{\bar D} \right\}$$

$$\times \left\{ \mathcal{G}[x',y']\langle\langle \delta_{\boldsymbol{\xi}\boldsymbol{\xi}'}\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}]\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\bar D}\rangle_{\bar D} + \langle\langle \overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}\mathcal{G}[\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\frac{\boldsymbol{\xi}\cdot\boldsymbol{\xi}'}{N}\delta[x'-\boldsymbol{J}\cdot\boldsymbol{\xi}']\delta[y'-\boldsymbol{B}\cdot\boldsymbol{\xi}']\rangle_{\bar D}\rangle_{\bar D} \right\} \Big\rangle_{\rm QRP;t}$$

$$= \eta^2 \frac{\partial^2}{\partial x\partial x'} \lim_{N\to\infty} \left\langle \left\{ \mathcal{O}(N^{-1}) + \mathcal{O}(N^{-\frac{1}{2}}) \right\} \left\{ \mathcal{O}(N^{-1}) + \mathcal{O}(N^{-\frac{1}{2}}) \right\} \right\rangle_{\rm QRP;t} = 0$$

For batch learning all diffusion matrix elements of (19) vanish in a straightforward manner. For on-line learning all diffusion terms vanish provided we can prove that the function $\mathcal{C}[\ldots]$ of (24) is zero. This is indeed the case within the present theory, as will be verified in the Appendix B.

B Replica Calculation of the Green's Function

The main objective of this Appendix is to calculate the Green's function $\mathcal{A}[\ldots]$, with which we obtain our macroscopic dynamic equations in explicit form. We first carry out the disorder averages, leading to an effective single-spin problem. The integrations are done by steepest descent, giving a saddle-point problem for replicated order parameters at each time step. In the saddle point equations we then make the replica symmetry (RS) ansatz, so that the limit $n \to 0$ can be taken. In addition we show that the two functions $\mathcal{B}[\ldots]$ and $\mathcal{C}[\ldots]$ do indeed vanish, as claimed.

B.1 Disorder Averaging

The fundamental quantities $\mathcal{A}[x,y;x',y']$, $\mathcal{B}[x,y;x',y']$, $\mathcal{C}[x,y;x',y';x'',y'']$, and $P[x,y]$, which control the macroscopic equations can be written as

$$\left\{\begin{array}{c} P[x,y] \\ \mathcal{A}[x,y;x',y'] \\ \mathcal{B}[x,y;x',y'] \\ \mathcal{C}[x,y;x',y';x'',y''] \end{array}\right\} = \lim_{\substack{N\to\infty \\ n\to 0}} \left\langle \left\langle\!\!\left\langle \right.\right.\right.$$

$$\int \prod_{\alpha} \left\{ \delta\left[N-(\boldsymbol{\sigma}^{\alpha})^2\right] \delta\left[\frac{NR}{\sqrt{Q}} - \boldsymbol{\tau}\cdot\boldsymbol{\sigma}^{\alpha}\right] d\boldsymbol{\sigma}^{\alpha} \prod_{x_\alpha y_\alpha} \delta\left[P[x_\alpha,y_\alpha] - P[x_\alpha,y_\alpha;\frac{\sqrt{Q}\boldsymbol{\sigma}^{\alpha}}{\sqrt{N}}]\right] \right\} \delta\left[x - \frac{\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right] \delta\left[y - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]$$

$$\left\{\begin{array}{c} 1 \\ (\boldsymbol{\xi}'\cdot\boldsymbol{\xi})\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}\delta\left[x' - \frac{\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \delta\left[y' - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \\ \left[\frac{1}{N}\sum_{i\neq j}\xi_i\xi_j\xi_i'\xi_j'\right]\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}\delta\left[x' - \frac{\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \delta\left[y' - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \\ \frac{1}{N}(\boldsymbol{\xi}\cdot\boldsymbol{\xi}'')(\boldsymbol{\xi}'\cdot\boldsymbol{\xi}'')\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}''}\overline{\delta}_{\boldsymbol{\xi}'\boldsymbol{\xi}''}\delta\left[x' - \frac{\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \delta\left[y' - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \delta\left[x'' - \frac{\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}''}{\sqrt{N}}\right] \delta\left[y'' - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}''}{\sqrt{N}}\right] \end{array}\right\} \Big\rangle_{\bar{D}} \Big\rangle_{\bar{D}} \Big\rangle_{\bar{D}} \Big\rangle_{\Xi}$$

We next use the definition of $P[x,y;\boldsymbol{J}]$, introduce integral representations for the δ-distributions involving $P[x,y]$, and obtain

$$\left\{\begin{array}{c} P[x,y] \\ \mathcal{A}[x,y;x',y'] \\ \mathcal{B}[x,y;x',y'] \\ \mathcal{C}[x,y;x',y';x'',y''] \end{array}\right\} = \lim_{\substack{N\to\infty \\ n\to 0}} \int \prod_{\alpha} \left\{ \delta\left[N-(\boldsymbol{\sigma}^{\alpha})^2\right] \delta\left[\frac{NR}{\sqrt{Q}} - \boldsymbol{\tau}\cdot\boldsymbol{\sigma}^{\alpha}\right] d\boldsymbol{\sigma}^{\alpha} \prod_{x_\alpha y_\alpha} e^{iN\pi_\alpha[x_\alpha,y_\alpha]P[x_\alpha,y_\alpha]} d\pi(x_\alpha,y_\alpha) \right\}$$

$$\times \left\langle \left\langle\!\!\left\langle e^{-iN\sum_\alpha \sum_{x_\alpha y_\alpha} \pi_\alpha[x_\alpha,y_\alpha]\langle \delta[x_\alpha - \frac{\sqrt{Q}\boldsymbol{\sigma}^{\alpha}\cdot\boldsymbol{\xi}'''}{\sqrt{N}}]\delta[y_\alpha - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}'''}{\sqrt{N}}]\rangle_{\bar{D}}} \right.\right.\right. \delta\left[x - \frac{\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right] \delta\left[y - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}\right]$$

$$\times \left\{\begin{array}{c} 1 \\ (\boldsymbol{\xi}'\cdot\boldsymbol{\xi})\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}\delta\left[x' - \frac{\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \delta\left[y' - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \\ \left[\frac{1}{N}\sum_{i\neq j}\xi_i\xi_j\xi_i'\xi_j'\right]\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}'}\delta\left[x' - \frac{\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \delta\left[y' - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \\ \frac{1}{N}(\boldsymbol{\xi}\cdot\boldsymbol{\xi}'')(\boldsymbol{\xi}'\cdot\boldsymbol{\xi}'')\overline{\delta}_{\boldsymbol{\xi}\boldsymbol{\xi}''}\overline{\delta}_{\boldsymbol{\xi}'\boldsymbol{\xi}''}\delta\left[x' - \frac{\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \delta\left[y' - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}'}{\sqrt{N}}\right] \delta\left[x'' - \frac{\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}''}{\sqrt{N}}\right] \delta\left[y'' - \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}''}{\sqrt{N}}\right] \end{array}\right\} \Big\rangle_{\bar{D}} \Big\rangle_{\bar{D}} \Big\rangle_{\bar{D}} \Big\rangle_{\Xi}$$

The summations involving (x_α, y_α) automatically lead to integrals, which can be performed due to the δ-distributions involved. We define new conjugate functions $\hat{P}_\alpha[x,y]$ via

$$\sum_{x_\alpha y_\alpha} \pi_\alpha[x_\alpha,y_\alpha] f[x_\alpha,y_\alpha] \;\to\; \int dx''dy''\; \hat{P}_\alpha[x'',y''] f[x'',y'']$$

We write averages over the training set explicitly in terms of the $p=\alpha N$ constituent vectors $\{\boldsymbol{\xi}^\mu\}$. Finally we introduce integrals representations for the remaining delta-distributions, and obtain the following expressions

(at this stage we will have to separate the various structurally different cases):

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

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

$$\left\langle e^{-\frac{i}{\alpha}\sum_\alpha\sum_\lambda \hat{P}_\alpha(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}, \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}) - i[\hat{x}\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}^\mu + \hat{y}\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\mu]/\sqrt{N}} \right\rangle_{\boldsymbol{\Xi}} \tag{103}$$

$$\left\{ \begin{array}{c} \mathcal{A}[x,y;x',y'] \\ \mathcal{B}[x,y;x',y'] \end{array} \right\} = \int \frac{d\hat{x}d\hat{x}'d\hat{y}d\hat{y}'}{(2\pi)^4} e^{i[x\hat{x}+x'\hat{x}'+y\hat{y}+y\hat{y}']} \lim_{\substack{N\to\infty \\ n\to 0}}$$

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

$$\left\langle \left\{ \begin{array}{c} (\boldsymbol{\xi}^\mu\cdot\boldsymbol{\xi}^\nu) \\ \frac{1}{N}\sum_{i\neq j}\xi_i^\mu\xi_i^\nu\xi_j^\mu\xi_j^\nu \end{array} \right\} e^{-\frac{i}{\alpha}\sum_\alpha\sum_\lambda \hat{P}_\alpha[\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}, \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}] - i[\hat{x}\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}^\mu + \hat{y}\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\mu + \hat{x}'\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}^\nu + \hat{y}'\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\nu]/\sqrt{N}} \right\rangle_{\boldsymbol{\Xi}} \tag{104}$$

$$\mathcal{C}[x,y;x',y';x'',y''] = \int \frac{d\hat{x}d\hat{x}'d\hat{x}''d\hat{y}d\hat{y}'d\hat{y}''}{(2\pi)^6} e^{i[x\hat{x}+x'\hat{x}'+x''\hat{x}''+y\hat{y}+y'\hat{y}'+y''\hat{y}'']} \lim_{\substack{N\to\infty \\ n\to 0}}$$

$$\int \prod_\alpha \left\{ \delta\left[N-(\boldsymbol{\sigma}^\alpha)^2\right] \delta\left[\frac{NR}{\sqrt{Q}} - \boldsymbol{\tau}\cdot\boldsymbol{\sigma}^\alpha\right] d\boldsymbol{\sigma}^\alpha e^{iN\int dx''dy''\ \hat{P}_\alpha[x'',y'']P[x'',y'']} \prod_{x''y''} d\hat{P}_\alpha(x'',y'') \right\} \frac{1}{p^3}\sum_{\mu\nu\rho=1}^p \overline{\delta}_{\mu\rho}\overline{\delta}_{\nu\rho}$$

$$\left\langle (\boldsymbol{\xi}^\mu\cdot\boldsymbol{\xi}^\rho)(\boldsymbol{\xi}^\nu\cdot\boldsymbol{\xi}^\rho) e^{-\frac{i}{\alpha}\sum_\alpha\sum_\lambda \hat{P}_\alpha[\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}, \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\lambda}{\sqrt{N}}] - i[\hat{x}\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}^\mu + \hat{y}\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\mu + \hat{x}'\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}^\nu + \hat{y}'\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\nu + \hat{x}''\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}^\rho + \hat{y}''\boldsymbol{\tau}\cdot\boldsymbol{\xi}^\rho]/\sqrt{N}} \right\rangle_{\boldsymbol{\Xi}} \tag{105}$$

The averages over the training sets $\langle \ldots \rangle_{\boldsymbol{\Xi}}$ in (103,104,105) will now be done separately. First we define some relevant objects:

$$\mathcal{D}[u,v] = \left\langle e^{-\frac{i}{\alpha}\sum_\alpha \hat{P}_\alpha(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}}{\sqrt{N}}, \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}) - i[u\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi} + v\boldsymbol{\tau}\cdot\boldsymbol{\xi}]/\sqrt{N}} \right\rangle_{\boldsymbol{\xi}} \tag{106}$$

$$\mathcal{E}_j[u,v] = \left\langle \sqrt{N}\xi_j\ e^{-\frac{i}{\alpha}\sum_\alpha \hat{P}_\alpha(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}}{\sqrt{N}}, \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}) - i[u\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi} + v\boldsymbol{\tau}\cdot\boldsymbol{\xi}]/\sqrt{N}} \right\rangle_{\boldsymbol{\xi}} \tag{107}$$

$$\mathcal{E}_{ij}[u,v] = \left\langle N\xi_i\xi_j\ e^{-\frac{i}{\alpha}\sum_\alpha \hat{P}_\alpha(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}}{\sqrt{N}}, \frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}}) - i[u\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi} + v\boldsymbol{\tau}\cdot\boldsymbol{\xi}]/\sqrt{N}} \right\rangle_{\boldsymbol{\xi}} \quad (i\neq j) \tag{108}$$

As we will see, all are of order $\mathcal{O}(N^0)$ as $N\to\infty$. We next use the permutation invariance of our integrations and summations with respect to pattern labels. First we calculate the first training sets average occurring in (103):

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

$$= e^{p\log\mathcal{D}[0,0]}\ \frac{\mathcal{D}[\hat{x},\hat{y}]}{\mathcal{D}[0,0]} \tag{109}$$

The prefactor $e^{p\log\mathcal{D}[0,0]}$, will turn out to take care of appropriate normalisation, and will drop out of the final result for all four functions $P[x,y]$, $\mathcal{A}[x,y;x',y']$, $\mathcal{B}[x,y;x',y']$ and $\mathcal{C}[x,y;x',y';x'',y'']$. Secondly we evaluate the training sets average of the expression for $\mathcal{A}[\ldots]$ in (104):

$$\begin{aligned}
\frac{1}{p^2}\sum_{\mu\neq\nu}^{p}\left\langle(\boldsymbol{\xi}^{\mu}\cdot\boldsymbol{\xi}^{\nu})\ e^{\cdots}\right\rangle_{\boldsymbol{\Xi}} &= \frac{p-1}{p}\left\langle(\boldsymbol{\xi}^{1}\cdot\boldsymbol{\xi}^{2})\ e^{\cdots}\right\rangle_{\boldsymbol{\Xi}} \\
&= \frac{p-1}{p}\sum_{j}\left\langle e^{-\frac{i}{\alpha}\sum_{\alpha}\hat{P}_{\alpha}(\frac{\sqrt{Q}\boldsymbol{\sigma}^{\alpha}\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}})}\right\rangle_{\boldsymbol{\xi}}^{p-2}\left\langle\xi_j e^{-\frac{i}{\alpha}\sum_{\alpha}\hat{P}_{\alpha}(\frac{\sqrt{Q}\boldsymbol{\sigma}^{\alpha}\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}})-i[\hat{x}\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}+\hat{y}\boldsymbol{\tau}\cdot\boldsymbol{\xi}]/\sqrt{N}}\right\rangle_{\boldsymbol{\xi}} \\
&\qquad\times\left\langle\xi_j e^{-\frac{i}{\alpha}\sum_{\alpha}\hat{P}_{\alpha}(\frac{\sqrt{Q}\boldsymbol{\sigma}^{\alpha}\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}})-i[\hat{x}'\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}+\hat{y}'\boldsymbol{\tau}\cdot\boldsymbol{\xi}]/\sqrt{N}}\right\rangle_{\boldsymbol{\xi}} \\
&= e^{p\log\mathcal{D}[0,0]}\left\{\frac{1}{N}\sum_{j=1}^{N}\frac{\mathcal{E}_j[\hat{x},\hat{y}]\mathcal{E}_j[\hat{x}',\hat{y}']}{\mathcal{D}^2[0,0]}+\mathcal{O}(N^{-1})\right\}
\end{aligned} \tag{110}$$

(provided we indeed show that $\mathcal{E}_j[u,v]=\mathcal{O}(N^0)$ as $N\to\infty$). Secondly, the training sets average of the expression for $\mathcal{B}[\ldots]$ in (104) is given by:

$$\begin{aligned}
\frac{1}{p^2}\sum_{\mu\neq\nu}^{p}\left\langle\frac{1}{N}\sum_{i\neq j}\xi_i^{\mu}\xi_i^{\nu}\xi_j^{\mu}\xi_j^{\nu}e^{\cdots}\right\rangle_{\boldsymbol{\Xi}} &= \frac{p-1}{pN}\sum_{i\neq j}\left\langle(\xi_i^1\xi_j^1)(\xi_i^2\xi_j^2)\ e^{\cdots}\right\rangle_{\boldsymbol{\Xi}} \\
&= \frac{p-1}{pN}\sum_{i\neq j}\left\langle e^{-\frac{i}{\alpha}\sum_{\alpha}\hat{P}_{\alpha}(\frac{\sqrt{Q}\boldsymbol{\sigma}^{\alpha}\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}})}\right\rangle_{\boldsymbol{\xi}}^{p-2}\left\langle\xi_i\xi_j e^{-\frac{i}{\alpha}\sum_{\alpha}\hat{P}_{\alpha}(\frac{\sqrt{Q}\boldsymbol{\sigma}^{\alpha}\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}})-i[\hat{x}\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}+\hat{y}\boldsymbol{\tau}\cdot\boldsymbol{\xi}]/\sqrt{N}}\right\rangle_{\boldsymbol{\xi}} \\
&\qquad\times\left\langle\xi_i\xi_j e^{-\frac{i}{\alpha}\sum_{\alpha}\hat{P}_{\alpha}(\frac{\sqrt{Q}\boldsymbol{\sigma}^{\alpha}\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}})-i[\hat{x}'\sqrt{Q}\boldsymbol{\sigma}^{1}\cdot\boldsymbol{\xi}+\hat{y}'\boldsymbol{\tau}\cdot\boldsymbol{\xi}]/\sqrt{N}}\right\rangle_{\boldsymbol{\xi}} \\
&= e^{p\log\mathcal{D}[0,0]}\left\{\frac{1}{N^3}\sum_{i\neq j=1}^{N}\frac{\mathcal{E}_{ij}[\hat{x},\hat{y}]\mathcal{E}_{ij}[\hat{x}',\hat{y}']}{\mathcal{D}^2[0,0]}+\mathcal{O}(N^{-\frac{3}{2}})\right\} = e^{p\log\mathcal{D}[0,0]}\left\{\mathcal{O}(N^{-1})\right\}
\end{aligned} \tag{111}$$

(provided we indeed show that $\mathcal{E}_{ij}[u,v]=\mathcal{O}(N^0)$ as $N\to\infty$). Finally we also obtain for the training sets average in (105), in a similar fashion:

$$\begin{aligned}
\frac{1}{p^3}\sum_{\rho=1}^{p}\sum_{\mu,\nu\neq\rho}^{p}\left\langle\frac{1}{N}(\boldsymbol{\xi}^{\mu}\cdot\boldsymbol{\xi}^{\rho})(\boldsymbol{\xi}^{\nu}\cdot\boldsymbol{\xi}^{\rho})\ e^{\cdots}\right\rangle_{\boldsymbol{\Xi}} &= \frac{p-1}{p^2N}\sum_{ij}\left\langle\xi_i^1\xi_j^1\xi_i^2\xi_j^2\ e^{\cdots}\right\rangle_{\boldsymbol{\Xi}}+\frac{(p-1)(p-2)}{p^2N}\sum_{ij}\left\langle\xi_i^1\xi_j^2\xi_i^3\xi_j^3\ e^{\cdots}\right\rangle_{\boldsymbol{\Xi}} \\
&= \sum_{i\neq j}\left\langle\xi_i^1\xi_j^1\xi_i^2\xi_j^2\ e^{\cdots}\right\rangle_{\boldsymbol{\Xi}}.\mathcal{O}(N^{-2})+\left\langle e^{\cdots}\right\rangle_{\boldsymbol{\Xi}}.\mathcal{O}(N^{-1})+\sum_{i\neq j}\left\langle\xi_i^1\xi_j^2\xi_i^3\xi_j^3\ e^{\cdots}\right\rangle_{\boldsymbol{\Xi}}.\mathcal{O}(N^{-1})+\sum_{i}\left\langle\xi_i^1\xi_i^2\ e^{\cdots}\right\rangle_{\boldsymbol{\Xi}}.\mathcal{O}(N^{-1}) \\
&= \mathcal{D}[0,0]^p\left\{\sum_{i\neq j}\mathcal{D}[\hat{x}'',\hat{y}'']\mathcal{E}_{ij}[\hat{x},\hat{y}]\mathcal{E}_{ij}[\hat{x}',\hat{y}'].\mathcal{O}(N^{-4})+\mathcal{O}(N^{-1})+\sum_{i\neq j}\mathcal{E}_i[\hat{x},\hat{y}]\mathcal{E}_j[\hat{x}',\hat{y}']\mathcal{E}_{ij}[\hat{x}'',\hat{y}''].\mathcal{O}(N^{-3})\right. \\
&\qquad\left.+\sum_{i}\mathcal{D}[\hat{x}'',\hat{y}'']\mathcal{E}_j[\hat{x},\hat{y}]\mathcal{E}_j[\hat{x}',\hat{y}'].\mathcal{O}(N^{-2})\right\} \\
&= e^{p\log\mathcal{D}[0,0]}\left\{\mathcal{O}(N^{-1})\right\}
\end{aligned} \tag{112}$$

We now work out (107) and we show that it is of order N^0. This is achieved by separating in the exponent the terms with site label $i=j$ from those with site labels $i\neq j$, followed by expansion in powers of the (relatively small) $i=j$ terms, and will involve the following two functions:

$$\mathcal{F}_1^\alpha[u,v]=\left\langle \partial_x\hat{P}_\alpha(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}})\;e^{-\frac{i}{\alpha}\sum_\alpha\hat{P}_\alpha(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}})-i[u\sqrt{Q}\boldsymbol{\sigma}^1\cdot\boldsymbol{\xi}+v\boldsymbol{\tau}\cdot\boldsymbol{\xi}]/\sqrt{N}}\right\rangle_{\boldsymbol{\xi}} \tag{113}$$

$$\mathcal{F}_2^\alpha[u,v]=\left\langle \partial_y\hat{P}_\alpha(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}})\;e^{-\frac{i}{\alpha}\sum_\alpha\hat{P}_\alpha(\frac{\sqrt{Q}\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}}{\sqrt{N}},\frac{\boldsymbol{\tau}\cdot\boldsymbol{\xi}}{\sqrt{N}})-i[u\sqrt{Q}\boldsymbol{\sigma}\cdot\boldsymbol{\xi}+v\boldsymbol{\tau}\cdot\boldsymbol{\xi}]/\sqrt{N}}\right\rangle_{\boldsymbol{\xi}} \tag{114}$$

Note that there is no need to calculate the auxiliary functions (108); we only need to verify their magnitude to scale as $\mathcal{O}(N^0)$ for $N\to\infty$.

$$\mathcal{E}_j[u,v]=\left\langle \sqrt{N}\xi_j e^{-\frac{i}{\alpha}\sum_\alpha\hat{P}_\alpha(\frac{\sqrt{Q}}{\sqrt{N}}[\sum_{i\neq j}\sigma_i^\alpha\xi_i+\sigma_j^\alpha\xi_j],\frac{1}{\sqrt{N}}[\sum_{i\neq j}\tau_i\xi_i+\tau_j\xi_j])-i[\frac{u\sqrt{Q}}{\sqrt{N}}[\sum_{i\neq j}\sigma_i^1\xi_i+\sigma_j^1\xi_j]+\frac{v}{\sqrt{N}}[\sum_{i\neq j}\tau_i\xi_i+\tau_j\xi_j]]}\right\rangle_{\boldsymbol{\xi}}$$

$$=\left\langle \sqrt{N}\xi_j e^{-\frac{i}{\alpha}\sum_\alpha\hat{P}_\alpha(\frac{\sqrt{Q}}{\sqrt{N}}\sum_{i\neq j}\sigma_i^\alpha\xi_i,\frac{1}{\sqrt{N}}\sum_{i\neq j}\tau_i\xi_i)-i[\frac{u\sqrt{Q}}{\sqrt{N}}\sum_{i\neq j}\sigma_i^1\xi_i+\frac{v}{\sqrt{N}}\sum_{i\neq j}\tau_i\xi_i]}\times\right.$$

$$\left. e^{-\frac{i\sqrt{Q}}{\alpha\sqrt{N}}\sum_\alpha\sigma_j^\alpha\xi_j\partial_x\hat{P}_\alpha(\frac{\sqrt{Q}}{\sqrt{N}}\sum_{i\neq j}\sigma_i^\alpha\xi_i,\frac{1}{\sqrt{N}}\sum_{i\neq j}\tau_i\xi_i)-\frac{i}{\alpha\sqrt{N}}\tau_j\xi_j\sum_\alpha\partial_y\hat{P}_\alpha(\frac{\sqrt{Q}}{\sqrt{N}}\sum_{i\neq j}\sigma_i^\alpha\xi_i,\frac{1}{\sqrt{N}}\sum_{i\neq j}\tau_i\xi_i)}\right.$$

$$\left.\times e^{-i[\frac{u\sqrt{Q}}{\sqrt{N}}\sigma_j^1\xi_j+\frac{v}{\sqrt{N}}\tau_j\xi_j]+\mathcal{O}(N^{-1})}\right\rangle_{\boldsymbol{\xi}}$$

$$=\left\langle e^{-\frac{i}{\alpha}\sum_\alpha\hat{P}_\alpha(\frac{\sqrt{Q}}{\sqrt{N}}\sum_{i\neq j}\sigma_i^\alpha\xi_i,\frac{1}{\sqrt{N}}\sum_{i\neq j}\tau_i\xi_i)-i[\frac{u\sqrt{Q}}{\sqrt{N}}\sum_{i\neq j}\sigma_i^1\xi_i+\frac{v}{\sqrt{N}}\sum_{i\neq j}\tau_i\xi_i]}\times\right.$$

$$\frac{1}{i}\left\{\frac{\sqrt{Q}}{\alpha}\sum_\alpha\sigma_j^\alpha\partial_x\hat{P}_\alpha(\frac{\sqrt{Q}}{\sqrt{N}}\sum_{i\neq j}\sigma_i^\alpha\xi_i,\frac{1}{\sqrt{N}}\sum_{i\neq j}\tau_i\xi_i)+\frac{1}{\alpha}\tau_j\sum_\alpha\partial_y\hat{P}_\alpha(\frac{\sqrt{Q}}{\sqrt{N}}\sum_{i\neq j}\sigma_i^\alpha\xi_i,\frac{1}{\sqrt{N}}\sum_{i\neq j}\tau_i\xi_i)\right.$$

$$\left.\left.+u\sqrt{Q}\sigma_j^1+v\tau_j+\mathcal{O}(N^{-\frac{1}{2}})\right\}\right\rangle_{\boldsymbol{\xi}}$$

so that

$$\mathcal{E}_j[u,v]=-iu\sqrt{Q}\sigma_j^1\mathcal{D}[u,v]-iv\tau_j\mathcal{D}[u,v]-\frac{i}{\alpha}\sqrt{Q}\sum_\alpha\sigma_j^\alpha\mathcal{F}_1^\alpha[u,v]-\frac{i}{\alpha}\tau_j\sum_\alpha\mathcal{F}_2^\alpha[u,v]+\mathcal{O}(N^{-\frac{1}{2}})$$

$$=-i\sqrt{Q}\sum_\alpha\sigma_j^\alpha\left[\frac{1}{\alpha}\mathcal{F}_1^\alpha[u,v]+u\delta_{\alpha1}\mathcal{D}[u,v]\right]-i\tau_j\sum_\alpha\left[\frac{1}{\alpha}\mathcal{F}_2^\alpha[u,v]+v\delta_{\alpha1}\mathcal{D}[u,v]\right]+\mathcal{O}(N^{-\frac{1}{2}}) \tag{115}$$

Repetition/extension of this argument, by separating in the exponent terms with two special indices (i,j) rather than one, and by subsequent expansion (whereby each term brings down a factor $N^{-\frac{1}{2}}$), immediately shows that terms of the form $\langle N\xi_i\xi_j\;e^{\cdots}\rangle_{\boldsymbol{\xi}}$ with $i\neq j$ will be of order $\mathcal{O}(N^0)$. This confirms that $\mathcal{E}_{ij}[u,v]=\mathcal{O}(N^0)$ and that (108) indeed scales as indicated. Note that the relevant combination of intensive terms in (110) can be abbreviated as $\mathcal{L}[u,v;u',v']=\frac{1}{N}\sum_j\mathcal{E}_j[u,v]\mathcal{E}_j[u',v']$:

$$\mathcal{L}[u,v;u',v']=-Q\sum_{\alpha\beta}q_{\alpha\beta}(\{\boldsymbol{\sigma}\})\left[\frac{1}{\alpha}\mathcal{F}_1^\alpha[u,v]+u\delta_{\alpha1}\mathcal{D}[u,v]\right]\left[\frac{1}{\alpha}\mathcal{F}_1^\beta[u',v']+u'\delta_{\beta1}\mathcal{D}[u',v']\right]$$

$$-R\sum_{\alpha\beta}\left[\frac{1}{\alpha}\mathcal{F}_1^\alpha[u,v]+u\delta_{\alpha1}\mathcal{D}[u,v]\right]\left[\frac{1}{\alpha}\mathcal{F}_2^\beta[u',v']+v'\delta_{\beta1}\mathcal{D}[u',v']\right]$$

$$-R\sum_{\alpha\beta}\left[\frac{1}{\alpha}\mathcal{F}_1^\alpha[u',v']+u'\delta_{\alpha 1}\mathcal{D}[u',v']\right]\left[\frac{1}{\alpha}\mathcal{F}_2^\beta[u,v]+v\delta_{\beta 1}\mathcal{D}[u,v]\right]$$

$$-\sum_{\alpha\beta}\left[\frac{1}{\alpha}\mathcal{F}_2^\alpha[u,v]+v\delta_{\alpha 1}\mathcal{D}[u,v]\right]\left[\frac{1}{\alpha}\mathcal{F}_2^\beta[u',v']+v'\delta_{\beta 1}\mathcal{D}[u',v']\right]+\mathcal{O}(N^{-\frac{1}{2}}) \tag{116}$$

where we have used the built-in properties $\frac{1}{N}\boldsymbol{\tau}\cdot\boldsymbol{\sigma}^\alpha=R/\sqrt{Q}$ and $\boldsymbol{\tau}^2=N$, and in which we find the spin-glass order parameters

$$q_{\alpha\beta}(\{\boldsymbol{\sigma}\})=\frac{1}{N}\sum_i\sigma_i^\alpha\sigma_i^\beta \tag{117}$$

Let us finally work out further the remaining fundamental objects $\mathcal{D}[\ldots]$ and $\mathcal{F}_{1,2}^\alpha[\ldots]$. The basic property to be used is that for large N the $n+1$ quantities $\{x_\alpha=\boldsymbol{\sigma}^\alpha\cdot\boldsymbol{\xi}/\sqrt{N},\ y=\boldsymbol{\tau}\cdot\boldsymbol{\xi}/\sqrt{N}\}$ inside averages of the form $\langle\ldots\rangle_{\boldsymbol{\xi}}$ will become (zero average but correlated) Gaussian variables, with probability distribution

$$P(x_1,\ldots,x_n,y)=\frac{\det^{\frac{1}{2}}\boldsymbol{A}}{(2\pi)^{(n+1)/2}}\ e^{-\frac{1}{2}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix}\cdot\boldsymbol{A}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix}}\qquad \boldsymbol{A}^{-1}=\begin{pmatrix}q_{11} & \cdots & q_{1n} & R/\sqrt{Q}\\ \vdots & & \vdots & \vdots\\ q_{n1} & \cdots & q_{nn} & R/\sqrt{Q}\\ R/\sqrt{Q} & \cdots & R/\sqrt{Q} & 1\end{pmatrix}$$

This allows us to write

$$\mathcal{D}[u,v]=\frac{\det^{\frac{1}{2}}\boldsymbol{A}}{(2\pi)^{(n+1)/2}}\int d\boldsymbol{x}dy\ e^{-\frac{1}{2}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix}\cdot\boldsymbol{A}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix}-\frac{i}{\alpha}\sum_\alpha\hat{P}_\alpha(\sqrt{Q}x_\alpha,y)-i[u\sqrt{Q}x_1+vy]} \tag{118}$$

$$\mathcal{F}_{1,2}^\alpha[u,v]=\frac{\det^{\frac{1}{2}}\boldsymbol{A}}{(2\pi)^{(n+1)/2}}\int d\boldsymbol{x}dy\ \partial_{1,2}\hat{P}_\alpha(\sqrt{Q}x_\alpha,y)\ e^{-\frac{1}{2}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix}\cdot\boldsymbol{A}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix}-\frac{i}{\alpha}\sum_\alpha\hat{P}_\alpha(\sqrt{Q}x_\alpha,y)-i[u\sqrt{Q}x_1+vy]} \tag{119}$$

Note that these quantities depend on the microscopic variables $\boldsymbol{\sigma}^\alpha$ only through the macroscopic observables $q_{\alpha\beta}(\{\boldsymbol{\sigma}\})$.

B.2 Derivation of Saddle-Point Equations

We will now combine the results (109,110,111,112) and (116) with the expressions (103,104,105). We use integral representations for the remaining delta functions, and isolate the observables $q_{\alpha\beta}$, by inserting

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

We hereby achieve a full factorisation over sites in the relevant quantities (note: the objects $\mathcal{D}[\ldots]$ and $\mathcal{L}[\ldots]$ depend on the microscopic variables only via $q_{\alpha\beta}(\{\boldsymbol{\sigma}\})$):

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

$$e^{iN[\sum_\alpha(\hat{Q}_\alpha+\hat{R}_\alpha R/\sqrt{Q})+\sum_{\alpha\beta}\hat{q}_{\alpha\beta}q_{\alpha\beta}+\sum_\alpha\int dx''dy''\ \hat{P}_\alpha(x'',y'')P[x'',y'']]+\alpha N\log\mathcal{D}[0,0]}$$

$$\prod_i \left\{ \int d\boldsymbol{\sigma}\ e^{-i\sum_\alpha [\hat{Q}_\alpha (\sigma_\alpha)^2 + \hat{R}_\alpha \tau_i \sigma_\alpha] - i\sum_{\alpha\beta} \hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta} \right\} \frac{\mathcal{L}[\hat{x},\hat{y};\hat{x}',\hat{y}']}{\mathcal{D}^2[0,0]}$$

and

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

$$e^{iN[\sum_\alpha (\hat{Q}_\alpha + \hat{R}_\alpha R/\sqrt{Q}) + \sum_{\alpha\beta}\hat{q}_{\alpha\beta}q_{\alpha\beta} + \sum_\alpha \int dx''dy''\ \hat{P}_\alpha(x'',y'')P[x'',y'']] + \alpha N \log \mathcal{D}[0,0]}$$

$$\prod_i \left\{ \int d\boldsymbol{\sigma}\ e^{-i\sum_\alpha [\hat{Q}_\alpha (\sigma_\alpha)^2 + \hat{R}_\alpha \tau_i \sigma_\alpha] - i\sum_{\alpha\beta} \hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta} \right\} \frac{\mathcal{D}[\hat{x},\hat{y}]}{\mathcal{D}[0,0]}$$

Both can be written in the form of an integral dominated by saddle-points:

$$\mathcal{A}[x,y;x',y'] = \int \frac{d\hat{x}d\hat{x}'d\hat{y}d\hat{y}'}{(2\pi)^4} e^{i[x\hat{x}+x'\hat{x}'+y\hat{y}+y\hat{y}']}$$

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

and

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

with

$$\Psi[\ldots] = i\sum_\alpha (\hat{Q}_\alpha + \hat{R}_\alpha R/\sqrt{Q}) + i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}q_{\alpha\beta} + i\sum_\alpha \int dx''dy''\ \hat{P}_\alpha(x'',y'')P[x'',y'']$$

$$+\alpha \log \mathcal{D}[0,0] + \lim_{N\to\infty}\frac{1}{N}\sum_i \log \int d\boldsymbol{\sigma}\ e^{-i\sum_\alpha [\hat{Q}_\alpha \sigma_\alpha^2 + \hat{R}_\alpha \tau_i \sigma_\alpha] - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta}$$

Finally we use that fact that the above expressions will be given by the intensive parts evaluated in the dominating saddle-point of Ψ. We can use the expression for $P[x,y]$ and its property $\int dxdy\ P[x,y] = 1$ to verify that all expressions are properly normalised (no overall prefactors are to be taken into account). We perform a simple transformation on some of our integration variables:

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

and finally we get

$$\mathcal{A}[x,y;x',y'] = \int \frac{d\hat{x}d\hat{x}'d\hat{y}d\hat{y}'}{(2\pi)^4} e^{i[x\hat{x}+x'\hat{x}'+y\hat{y}+y\hat{y}']} \lim_{n\to 0} \frac{\mathcal{L}[\hat{x},\hat{y};\hat{x}',\hat{y}']}{\mathcal{D}^2[0,0]} \tag{120}$$

$$P[x,y] = \int \frac{d\hat{x}d\hat{y}}{(2\pi)^2} e^{i[x\hat{x}+y\hat{y}]} \lim_{n\to 0} \frac{\mathcal{D}[\hat{x},\hat{y}]}{\mathcal{D}[0,0]} \tag{121}$$

in which all functions are to be evaluated upon choosing for the order parameters the appropriate saddle-points of Ψ (variation with respect to $\boldsymbol{q}$, $\hat{\boldsymbol{q}}$, $\hat{\boldsymbol{Q}}$, $\hat{\boldsymbol{R}}$ and $\{\hat{P}\}$), which itself takes the form:

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

$$+\alpha \log \mathcal{D}[0,0] + \lim_{N\to\infty}\frac{1}{N}\sum_i \log \int d\boldsymbol{\sigma}\ e^{-i\tau_i\sqrt{Q}\sum_\alpha \hat{R}_\alpha\sigma_\alpha - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta} \tag{122}$$

With $\mathcal{D}[\ldots]$ given by (118), which depends on the variational parameters $\{\hat{P}\}$ and $q_{\alpha\beta}$ only. The function $\mathcal{L}[\ldots]$ is given by (116). The 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 $(\boldsymbol{J}^a, \boldsymbol{J}^b)$, in two systems a and b with the same realization of the training set (see e.g. [16]):

$$\Big\langle P(q)\Big\rangle_{\Xi} = \Big\langle \Big\langle\!\Big\langle \delta\left[q - \frac{\boldsymbol{J}^a\cdot\boldsymbol{J}^b}{|\boldsymbol{J}^a||\boldsymbol{J}^b|}\right]\Big\rangle\!\Big\rangle \Big\rangle_{\Xi} = \lim_{n\to 0}\frac{1}{n(n-1)}\sum_{\alpha\neq\beta}\delta[q-q_{\alpha\beta}] \tag{123}$$

Note that upon applying the above procedure to the functions $\mathcal{B}[\ldots]$ and $\mathcal{C}[\ldots]$ in (104,105) we find again integrals dominated by the dominant saddle-point of Ψ; here, in view of (111) and (112), the intensive parts are zero, and thus

$$\mathcal{B}[x,y;x',y'] = \mathcal{C}[x,y;x',y';x'',y''] = 0 \tag{124}$$

as anticipated earlier.

B.3 Replica-Symmetric Saddle-Points

We now make the replica symmetric (RS) ansatz in the extremisation problem, which according to (123) is equivalent to assuming ergodicity. With a modest amount of foresight we put

$$q_{\alpha\beta} = q_0\delta_{\alpha\beta} + q[1-\delta_{\alpha\beta}], \qquad \hat{q}_{\alpha\beta} = \frac{1}{2}i[r-r_0\delta_{\alpha\beta}], \qquad \hat{R}_\alpha = i\rho, \qquad \hat{Q}_\alpha = i\phi, \qquad \hat{P}_\alpha[u,v] = i\chi[u,v]$$

This converts the quantity Ψ of equation (122) for small n into

$$\lim_{n\to 0}\frac{1}{n}\Psi[\ldots] = -\phi(1-q_0) - \rho R + \frac{1}{2}qr - \frac{1}{2}q_0(r-r_0) - \int dx''dy''\ \chi[x'',y'']P[x'',y'']$$

$$+\lim_{n\to 0}\frac{\alpha}{n}\log\mathcal{D}[0,0] + \lim_{n\to 0}\lim_{N\to\infty}\frac{1}{Nn}\sum_i \log\int Dz\int d\boldsymbol{\sigma}\ e^{\tau_i\rho\sqrt{Q}\sum_\alpha\sigma_\alpha - \frac{1}{2}r_0\sum_\alpha\sigma_\alpha^2 + z\sqrt{r}\sum_\alpha\sigma_\alpha}$$

with the abbreviation $Dz = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}z^2}dz$. We do the Gaussian integral in the last term, and expand the result for small n:

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

$$-\int dx''dy''\ \chi[x'',y'']P[x'',y''] + \lim_{n\to 0}\frac{\alpha}{n}\log\mathcal{D}[0,0] + \text{ const} \tag{125}$$

Note that 'const' refers to terms which do not depend on the order parameters to be varied, and will thus not show up in saddle-point equations; such terms can, however, depend on time via quantities such as (Q,R). At this stage it is useful to work out four of our saddle-point equations:

$$\frac{\partial\Psi}{\partial\phi} = \frac{\partial\Psi}{\partial r} = \frac{\partial\Psi}{\partial\rho} = \frac{\partial\Psi}{\partial r_0} = 0: \qquad q_0 = 1, \qquad r_0 = \frac{1}{1-q}, \qquad \rho = \frac{R}{Q(1-q)}, \qquad r = \frac{qQ-R^2}{Q(1-q)^2}$$

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

$$\lim_{n\to 0}\frac{1}{n}\Psi[q,\{\chi\}] = \frac{1-R^2/Q}{2(1-q)} + \frac{1}{2}\log(1-q) - \int dx'dy'\ \chi[x',y']P[x',y'] + \lim_{n\to 0}\frac{\alpha}{n}\log\mathcal{D}[0,0;q,\{\chi\}] + \text{ const} \tag{126}$$

Finally we have to work out the RS version of $\mathcal{D}[u,v;q,\{\chi\}]$:

$$\mathcal{D}[u,v;\chi,q,1] = \frac{\det^{\frac{1}{2}}\boldsymbol{A}}{(2\pi)^{(n+1)/2}}\int d\boldsymbol{x}dy\ e^{-\frac{1}{2}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix}\cdot\boldsymbol{A}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix} + \frac{1}{\alpha}\sum_\alpha\chi(\sqrt{Q}x_\alpha,y) - i[u\sqrt{Q}x_1+vy]} \tag{127}$$

with

$$\boldsymbol{A}^{-1} = \begin{pmatrix} 1 & \cdots & q & R/\sqrt{Q} \\ \vdots & & \vdots & \vdots \\ q & \cdots & 1 & R/\sqrt{Q} \\ R/\sqrt{Q} & \cdots & R/\sqrt{Q} & 1 \end{pmatrix}$$

The inverse of the above matrix is found to be

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

With this expression, and upon linearising the terms in the exponents which are quadratic in $\boldsymbol{x}$ in the usual manner with Gaussian integrals, we obtain

$$\mathcal{D}[u,v;q,\{\chi\}] = \frac{\int d\boldsymbol{x}dy\ e^{-\frac{1}{2}\boldsymbol{x}\cdot\boldsymbol{C}\boldsymbol{x}-\frac{1}{2}by^2-\gamma y\sum_{\alpha=1}^n x_\alpha+\frac{1}{\alpha}\sum_\alpha \chi[\sqrt{Q}x_\alpha,y]-i[u\sqrt{Q}x_1+vy]}}{\int d\boldsymbol{x}dy\ e^{-\frac{1}{2}\boldsymbol{x}\cdot\boldsymbol{C}\boldsymbol{x}-\frac{1}{2}by^2-\gamma y\sum_{\alpha=1}^n x_\alpha}}$$

$$= \frac{\int DzDy e^{-ivy/\sqrt{b}}\left[\int dx\ e^{-\frac{x^2}{2(1-q)}+[z\sqrt{d}-\gamma\frac{y}{\sqrt{b}}]x+\frac{1}{\alpha}\chi[\sqrt{Q}x,\frac{y}{\sqrt{b}}]}\right]^{n-1}\int dx\ e^{-\frac{x^2}{2(1-q)}+[z\sqrt{d}-\gamma\frac{y}{\sqrt{b}}]x+\frac{1}{\alpha}\chi[\sqrt{Q}x,\frac{y}{\sqrt{b}}]-iu\sqrt{Q}x}}{\int DzDy\left[\int dx\ e^{-\frac{1}{2(1-q)}x^2+[z\sqrt{d}-\gamma\frac{y}{\sqrt{b}}]x}\right]^n} \tag{128}$$

For the saddle-point problem we only need to calculate $\lim_{n\to 0}\frac{\alpha}{n}\log\mathcal{D}[0,0;q,\{\chi\}]$:

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

with γ and d evaluated in the limit $n\to 0$. Equivalently we can define

$$A = R/Q(1-q) \qquad B = \sqrt{qQ-R^2}/Q(1-q) \tag{129}$$

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 doing the x-integration in the denominator of this expression we can write the explicit expression for the surface Ψ to be extremised with respect to q and the function $\chi[x,y]$, apart from irrelevant constants, in the surprisingly simple form (with the short-hand (129)):

$$\lim_{n\to 0}\frac{1}{n}\Psi[q,\{\chi\}] = \frac{1-\alpha-R^2/Q}{2(1-q)} + \frac{1}{2}(1-\alpha)\log(1-q) - \int dxdy\ \chi[x,y]P[x,y]$$
$$+\ \alpha\int DzDy\ \log\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]} \tag{130}$$

Note that (130) is to be *minimised*, 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 the function $\chi[x,y]$ (obtained from the n-fold occurrence of the original function $\hat{P}$, multiplied by i, which also leads to curvature sign change).

The remaining saddle point equations are obtained by variation of (130) with respect to χ and q. Functional variation with respect to χ gives:

$$\text{for all } x,y: \qquad P[x,y]=\frac{e^{-\frac{1}{2}y^2}}{\sqrt{2\pi}}\int Dz\left\{\frac{e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}{\int dx'\ e^{-\frac{x'^2}{2Q(1-q)}+x'[Ay+Bz]+\frac{1}{\alpha}\chi[x',y]}}\right\} \tag{131}$$

Note that $P[x,y]=P[x|y]P[y]$ with $P[y]=(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}y^2}$, as could have been expected. Next we vary q, and use (131) wherever possible:

$$\frac{1-\alpha-R^2/Q}{2(1-q)^2}-\frac{1-\alpha}{2(1-q)}=\alpha\int DzDy\left\{\frac{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}\left[\frac{x^2}{2Q(1-q)^2}-x[y\frac{\partial A}{\partial q}+z\frac{\partial B}{\partial q}]\right]}{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}\right\}$$

giving

$$\begin{aligned}&\int dxdy\ P[x,y](x-Ry)^2+(R^2-qQ)(\frac{1}{\alpha}-1)\\&=\left[2\sqrt{qQ-R^2}+\frac{Q(1-q)}{\sqrt{qQ-R^2}}\right]\int DzDy\ z\left[\frac{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}x}{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}\right]\end{aligned} \tag{132}$$

B.4 Explicit Expression for the Green's Function

In order to work out the Green's function (120) we need the function $\mathcal{L}[u,v;u',v']$ as defined in (116) which, in turn, is given in terms of the integrals (118,119). First we calculate the $n\to 0$ limit of $D[u,v;q,\{\chi\}]$ (128), and simplify the result with the saddle-point equation (131):

$$\begin{aligned}\lim_{n\to 0}\mathcal{D}[u,v;q,\{\chi\}]&=\int DzDy\ e^{-ivy}\frac{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]-iux}}{\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}\\&=\int dxdyP[x,y]e^{-ivy-iux}\end{aligned} \tag{133}$$

Next we work out the quantities $F^{\alpha}_{1,2}[u,v]$ of equation (119) in RS ansatz, using Gaussian linearizations:

$$\lim_{n\to 0}\mathcal{F}^{\alpha}_{1,2}[u,v]=\ i\lim_{n\to 0}\ \frac{\int d\boldsymbol{x}dy\ \partial_{1,2}\chi[\sqrt{Q}x_\alpha,y]\ e^{-\frac{1}{2}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix}\cdot\boldsymbol{A}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix}+\frac{1}{\alpha}\sum_\alpha\chi[\sqrt{Q}x_\alpha,y]-i[u\sqrt{Q}x_1+vy]}}{\int d\boldsymbol{x}dy\ \ e^{-\frac{1}{2}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix}\cdot\boldsymbol{A}\begin{pmatrix}x_1\\ \vdots\\ x_n\\ y\end{pmatrix}}}$$

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

The replica permutation symmetries of this expression allow us to conclude

$$\lim_{n\to 0}\mathcal{F}^{\alpha}_{\lambda}[u,v]=\delta_{\alpha 1}F^1_\lambda[u,v]+(1-\delta_{\alpha 1})F^2_\lambda[u,v] \tag{134}$$

where

$$F^1_{1,2}[u,v] = i\int dxdy\ P[x,y]e^{-ivy-iux}\partial_{1,2}\chi[x,y] \tag{135}$$

$$F^2_{1,2}[u,v] = i\int DyDz\ e^{-ivy}\frac{\left[\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}\partial_{1,2}\chi[x,y]\right]\left[\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]-iux}\right]}{\left[\int dx\ e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}\right]^2} \tag{136}$$

We can now proceed to the calculation of (116). First we note that the basic building blocks of (116) are most easily expressed in terms of the functions

$$G_1[u,v] = \frac{1}{\alpha}\mathcal{F}^1_1[u,v] + u\mathcal{D}[u,v] \qquad \tilde{G}_1[u,v] = \frac{1}{\alpha}\mathcal{F}^2_1[u,v] \tag{137}$$

$$G_2[u,v] = \frac{1}{\alpha}\mathcal{F}^1_2[u,v] + v\mathcal{D}[u,v] \qquad \tilde{G}_2[u,v] = \frac{1}{\alpha}\mathcal{F}^2_2[u,v] \tag{138}$$

With these short-hands we obtain, upon performing the summations over replica indices in (116):

$$\begin{aligned}\mathcal{L}[u,v;u',v'] = &-Q(1-q)G_1[u,v]G_1[u',v'] - Q(1-q)(n-1)\tilde{G}_1[u,v]\tilde{G}_1[u',v']\\ &-Qq\left[G_1[u,v]+(n-1)\tilde{G}_1[u,v]\right]\left[G_1[u',v']+(n-1)\tilde{G}_1[u',v']\right]\\ &-R\left[G_1[u,v]+(n-1)\tilde{G}_1[u,v]\right]\left[G_2[u',v']+(n-1)\tilde{G}_2[u',v']\right]\\ &-R\left[G_1[u',v']+(n-1)\tilde{G}_1[u',v']\right]\left[G_2[u,v]+(n-1)\tilde{G}_2[u,v]\right]\\ &-\left[G_2[u,v]+(n-1)\tilde{G}_2[u,v]\right]\left[G_2[u',v']+(n-1)\tilde{G}_2[u',v']\right]\end{aligned}$$

and so

$$\begin{aligned}\lim_{n\to 0}\mathcal{L}[u,v;u',v'] = &-Q(1-q)\left[G_1[u,v]G_1[u',v'] - \tilde{G}_1[u,v]\tilde{G}_1[u',v']\right]\\ &-Qq\left[G_1[u,v]-\tilde{G}_1[u,v]\right]\left[G_1[u',v']-\tilde{G}_1[u',v']\right]\\ &-R\left[G_1[u,v]-\tilde{G}_1[u,v]\right]\left[G_2[u',v']-\tilde{G}_2[u',v']\right] - R\left[G_1[u',v']-\tilde{G}_1[u',v']\right]\left[G_2[u,v]-\tilde{G}_2[u,v]\right]\\ &-\left[G_2[u,v]-\tilde{G}_2[u,v]\right]\left[G_2[u',v']-\tilde{G}_2[u',v']\right]\end{aligned}$$

With the Fourier transforms of the functions $G[\ldots]$, given by

$$\hat{G}_1[\hat{u},\hat{v}] = \int\frac{dudv}{(2\pi)^2}e^{iu\hat{u}+iv\hat{v}}\left[\frac{1}{\alpha}\mathcal{F}^1_1[u,v]+u\mathcal{D}[u,v]\right] \qquad \overline{G}_1[\hat{u},\hat{v}] = \frac{1}{\alpha}\int\frac{dudv}{(2\pi)^2}e^{iu\hat{u}+iv\hat{v}}\mathcal{F}^2_1[u,v] \tag{139}$$

$$\hat{G}_2[\hat{u},\hat{v}] = \int\frac{dudv}{(2\pi)^2}e^{iu\hat{u}+iv\hat{v}}\left[\frac{1}{\alpha}\mathcal{F}^1_2[u,v]+v\mathcal{D}[u,v]\right] \qquad \overline{G}_2[\hat{u},\hat{v}] = \frac{1}{\alpha}\int\frac{dudv}{(2\pi)^2}e^{iu\hat{u}+iv\hat{v}}\mathcal{F}^2_2[u,v] \tag{140}$$

the Green's function $\mathcal{A}[x,y;x',y']$ (120) can now be written in explicit form as

$$\begin{aligned}\mathcal{A}[x,y;x',y'] = &-Q(1-q)\left[\hat{G}_1[x,y]\hat{G}_1[x',y'] - \overline{G}_1[x,y]\overline{G}_1[x',y']\right]\\ &-Qq\left[\hat{G}_1[x,y]-\overline{G}_1[x,y]\right]\left[\hat{G}_1[x',y']-\overline{G}_1[x',y']\right]\\ &-R\left[\hat{G}_1[x,y]-\overline{G}_1[x,y]\right]\left[\hat{G}_2[x',y']-\overline{G}_2[x',y']\right] - R\left[\hat{G}_1[x',y']-\overline{G}_1[x',y']\right]\left[\hat{G}_2[x,y]-\overline{G}_2[x,y]\right]\end{aligned}$$

$$-\left[\hat{G}_2[x,y]-\overline{G}_2[x,y]\right]\left[\hat{G}_2[x',y']-\overline{G}_2[x',y']\right] \tag{141}$$

Finally, working out the four relevant Fourier transforms, using (133,135,136), gives:

$$\hat{G}_1[x,y]=iP[x,y]\left[\frac{1}{\alpha}\frac{\partial}{\partial x}\chi[x,y]-\frac{\partial}{\partial x}\log P[x,y]\right] \tag{142}$$

$$\hat{G}_2[x,y]=iP[x,y]\left[\frac{1}{\alpha}\frac{\partial}{\partial y}\chi[x,y]-\frac{\partial}{\partial y}\log P[x,y]\right] \tag{143}$$

$$\overline{G}_1[x,y]=\frac{i}{\alpha}P[y]\int Dz\frac{\left[\int dx'\ e^{-\frac{x'^2}{2Q(1-q)}+x'[Ay+Bz]+\frac{1}{\alpha}\chi[x',y]}\partial_1\chi[x',y]\right]e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}{\left[\int dx'\ e^{-\frac{x'^2}{2Q(1-q)}+x'[Ay+Bz]+\frac{1}{\alpha}\chi[x',y]}\right]^2} \tag{144}$$

$$\overline{G}_2[x,y]=\frac{i}{\alpha}P[y]\int Dz\frac{\left[\int dx'\ e^{-\frac{x'^2}{2Q(1-q)}+x'[Ay+Bz]+\frac{1}{\alpha}\chi[x',y]}\partial_2\chi[x',y]\right]e^{-\frac{x^2}{2Q(1-q)}+x[Ay+Bz]+\frac{1}{\alpha}\chi[x,y]}}{\left[\int dx'\ e^{-\frac{x'^2}{2Q(1-q)}+x'[Ay+Bz]+\frac{1}{\alpha}\chi[x',y]}\right]^2} \tag{145}$$

with $P[y]=(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}y^2}$.

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:

$$M[x,y]=e^{-\frac{x^2}{2Q(1-q)}+Axy+\frac{1}{\alpha}\chi[x,y]}\qquad\qquad \langle f[x,y,z]\rangle_\star=\frac{\int dx\ M[x,y]e^{Bxz}f[x,y,z]}{\int dx\ M[x,y]e^{Bxz}} \tag{146}$$

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 (141) translates into the following expression:

$$\tilde{\mathcal{A}}[x,y;x',y']=Q(1-q)\left[J_1[x,y]J_1[x',y']-\tilde{J}_1[x,y]\tilde{J}_1[x',y']\right]+Qq\left[J_1[x,y]-\tilde{J}_1[x,y]\right]\left[J_1[x',y']-\tilde{J}_1[x',y']\right]$$
$$+R\left[J_1[x,y]-\tilde{J}_1[x,y]\right]J_2[x',y']+R\left[J_1[x',y']-\tilde{J}_1[x',y']\right]J_2[x,y]+J_2[x,y]J_2[x',y'] \tag{147}$$

with

$$J_1[X,Y]=\frac{\partial}{\partial X}\log\frac{M[X,Y]}{P[X|Y]}+\frac{X-RY}{Q(1-q)}$$

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

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

It turns out that significant simplification of the result (147) is possible, upon using the following two identities to rewrite the functions $J_1[\ldots]$, $\tilde{J}_1[\ldots]$ and $J_2[\ldots]$:

$$\langle\frac{\partial}{\partial x}\log M[x,y]\rangle_\star=-Bz \tag{148}$$

$$\langle\frac{\partial}{\partial y}\log M[x,y]\rangle_\star=\frac{\partial}{\partial y}\log\int dx\ e^{Bxz}M[x,y] \tag{149}$$

Identity (148) results upon integrating by parts with respect to x, whereas identity (149) is a direct consequence of y dependencies occurring in $M[x,y]$ only. Note that $B = \sqrt{qQ-R^2}/Q(1-q)$. To achieve the desired simplification of $\tilde{\mathcal{A}}[x,y;x',y']$ we define the following object:

$$\Phi[X,y] = \Big\{ Q(1-q)P[X|y] \Big\}^{-1} \int Dz \langle X-x \rangle_\star \langle \delta[X-x] \rangle_\star \tag{150}$$

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

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

$$J_2[X,Y] = Y - R\Phi[X,Y]$$

and consequently

$$\mathcal{A}[x,y;x',y'] = P[x,y]\tilde{\mathcal{A}}[x,y;x',y']P[x',y'] \tag{151}$$

$$\tilde{\mathcal{A}}[x,y;x',y'] = yy' + (x-Ry)\Phi[x',y'] + (x'-Ry')\Phi[x,y] - (Q-R^2)\Phi[x,y]\Phi[x',y'] \tag{152}$$

with $\Phi[x,y]$ as given in (150).

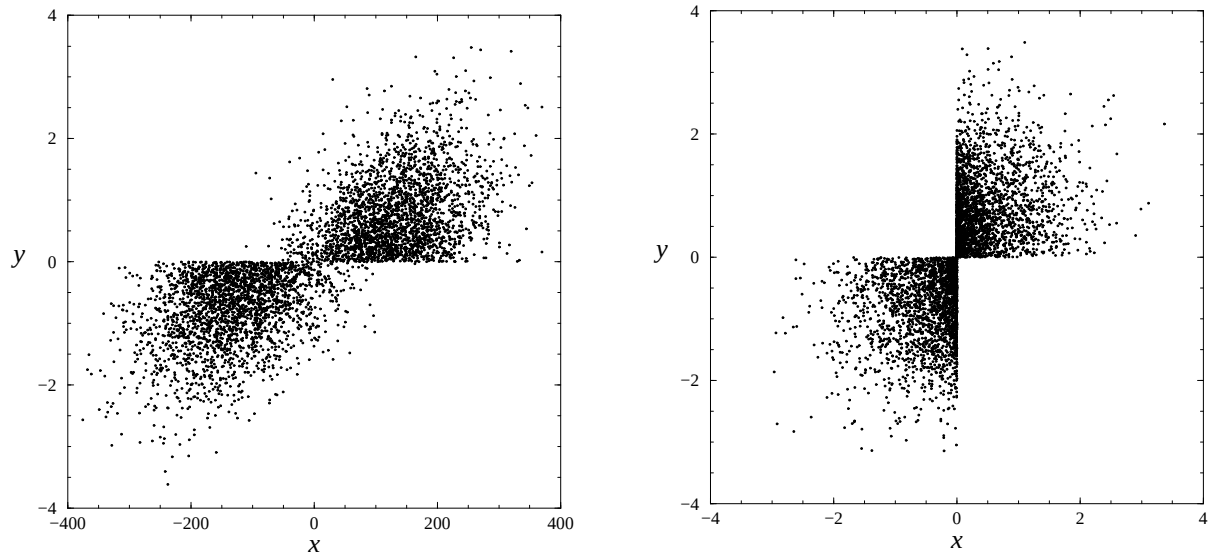

Figure 1: Student and teacher fields $(x, y) = (\boldsymbol{J}\cdot\boldsymbol{\xi}, \boldsymbol{B}\cdot\boldsymbol{\xi})$ as observed during numerical simulations of on-line learning (learning rate $\eta = 1$) in a perceptron of size $N = 10,000$ at $t = 50$, using 'questions' from a restricted training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

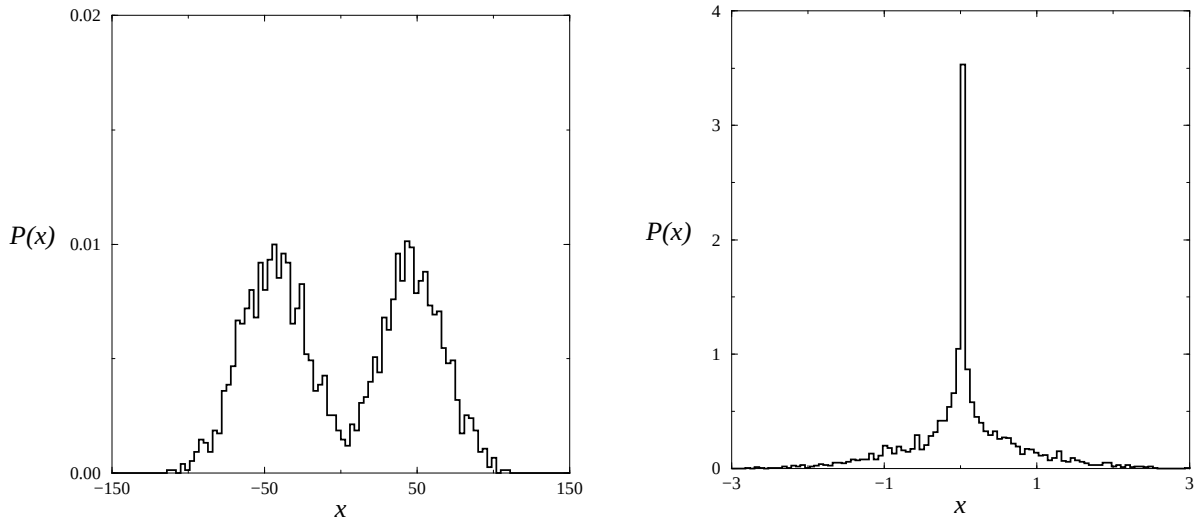

Figure 2: Distribution $P(x)$ of student fields as observed during numerical simulations of on-line learning (learning rate $\eta = 1$) in a perceptron of size $N = 10,000$, using 'questions' from a restricted training set of size $p = N/4$. Left: Hebbian learning, measured at $t = 10$. Right: AdaTron 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.

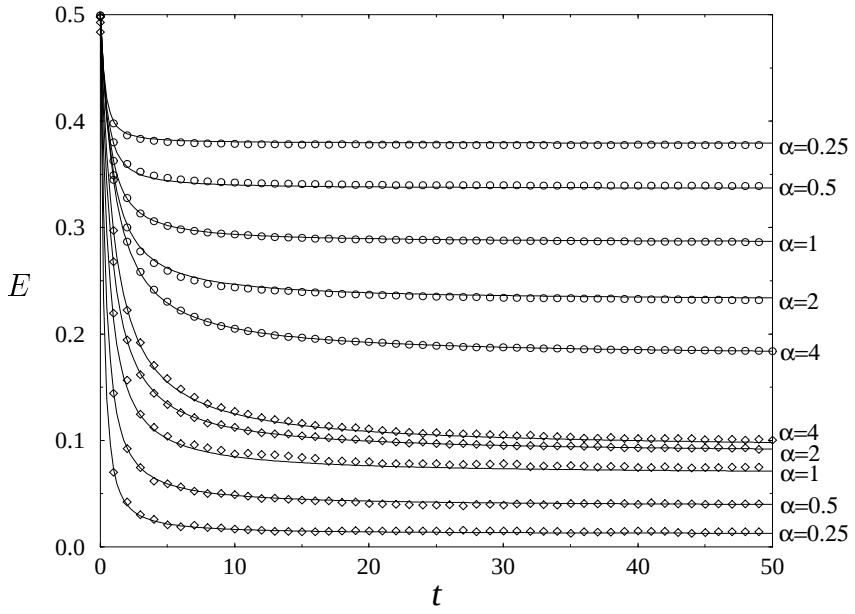

Figure 3: On-line Hebbian learning, simulations versus theoretical predictions, for $\eta = 1$ and $\alpha \in \{0.25, 0.5, 1.0, 2.0, 4.0\}$ ($N = 10,000$). Upper curves: generalization errors as functions of time. Lower curves: training errors as functions of time. Circles: simulation results for $E_{\rm g}$; diamonds: simulation results for $E_{\rm t}$. Solid lines: corresponding predictions of dynamical replica theory.

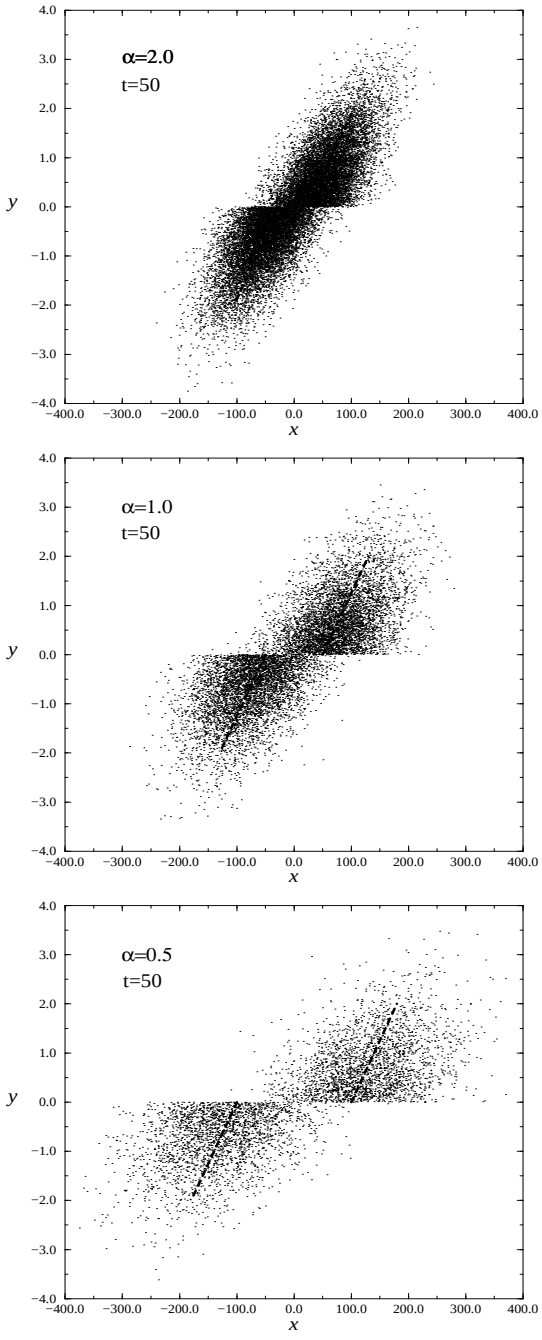

Figure 4: Comparison between simulation results for on-line Hebbian learning (system size $N = 10,000$) and dynamical replica theory, for $\eta = 1$ and $\alpha \in \{0.5, 1.0, 2.0\}$. Dots: local fields $(x, y) = (\boldsymbol{J}\cdot\boldsymbol{\xi}, \boldsymbol{B}\cdot\boldsymbol{\xi})$ (calculated 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, $\overline{x}(y) = Ry + (\eta t/\alpha)\ \mathrm{sgn}(y)$.

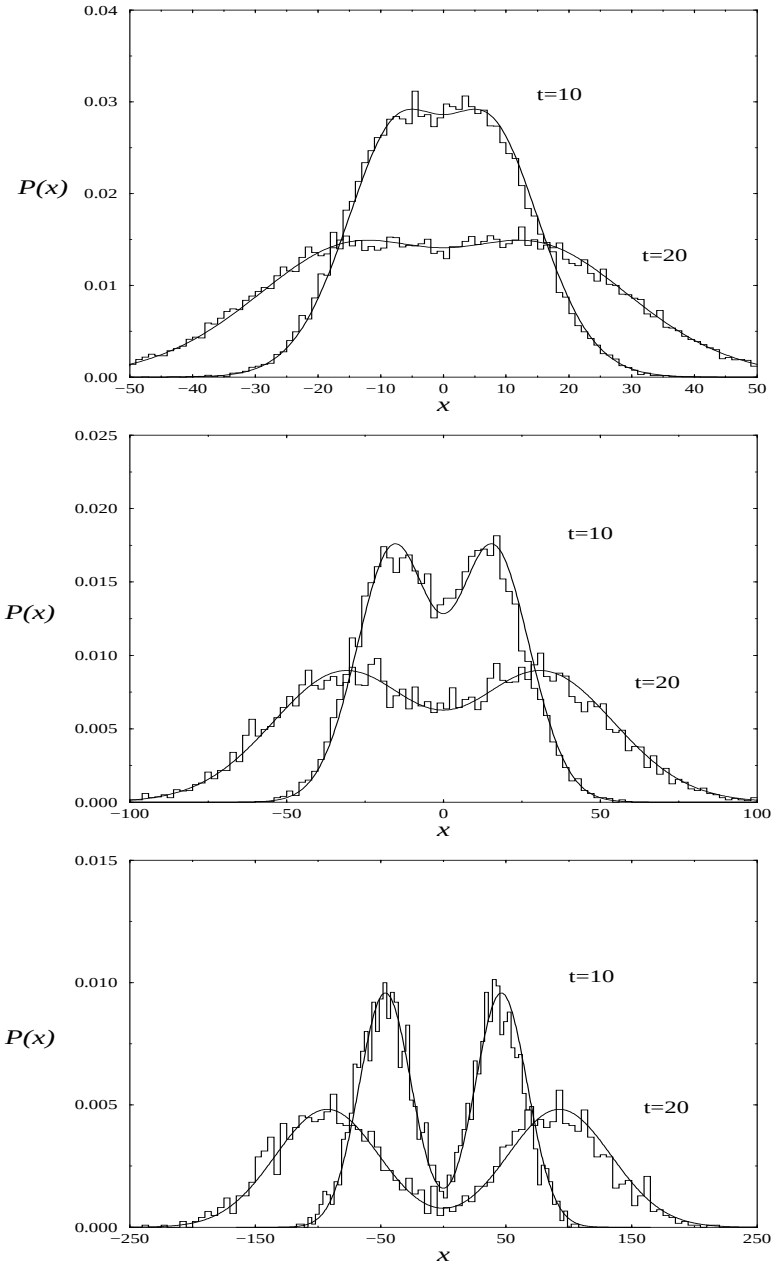

Figure 5: Simulations of on-line Hebbian learning with $\eta = 1$ and $N = 10,000$. Histograms: student field distributions measured at $t = 10$ and $t = 20$. Lines: theoretical predictions for student field distributions. $\alpha = 4$ (upper), $\alpha = 1$ (middle), $\alpha = 0.25$ (lower).

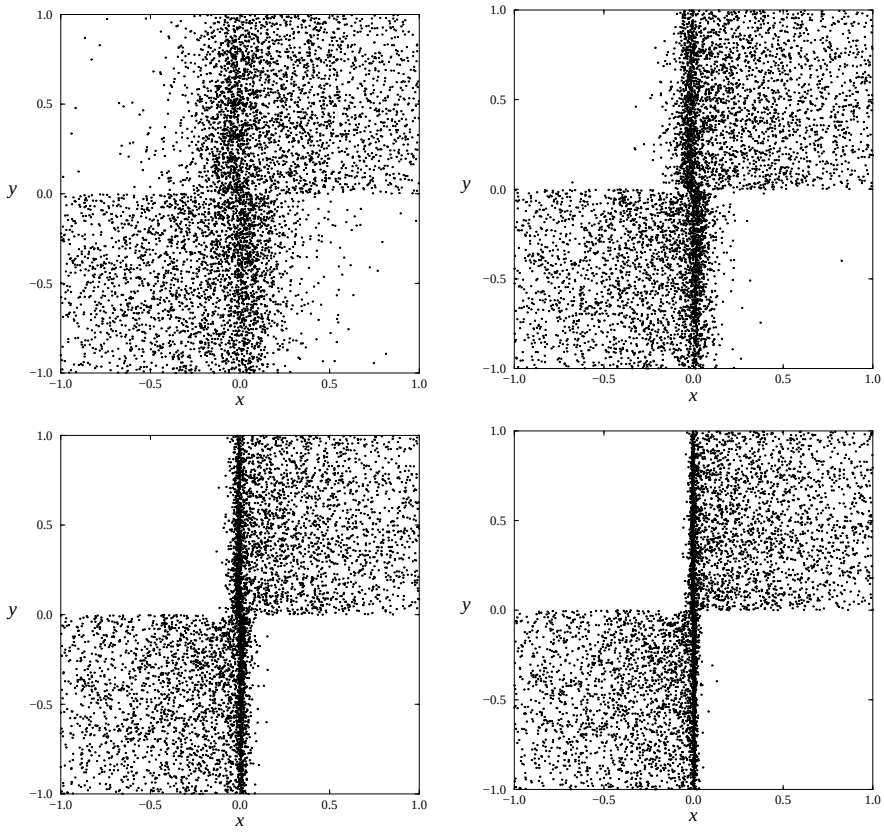

Figure 6: Numerical simulations of on-line Adatron learning, with $N = 10{,}000$, $\alpha = 1$ and $\eta = \frac{1}{2}$. The scatter plots show the observed student and teacher fields $(x, y) = (\boldsymbol{J} \cdot \boldsymbol{\xi}, \boldsymbol{B} \cdot \boldsymbol{\xi})$ at times $t = 5$ (upper left), $t = 10$ (upper right), $t = 15$ (lower left) and $t = 20$ (lower right), as measured during simulations for the data in the training set $\tilde{D}$, drawn as points in the (x, y) plane. Note the development over time of an increasingly narrow 'ridge' along the line $x = 0$.

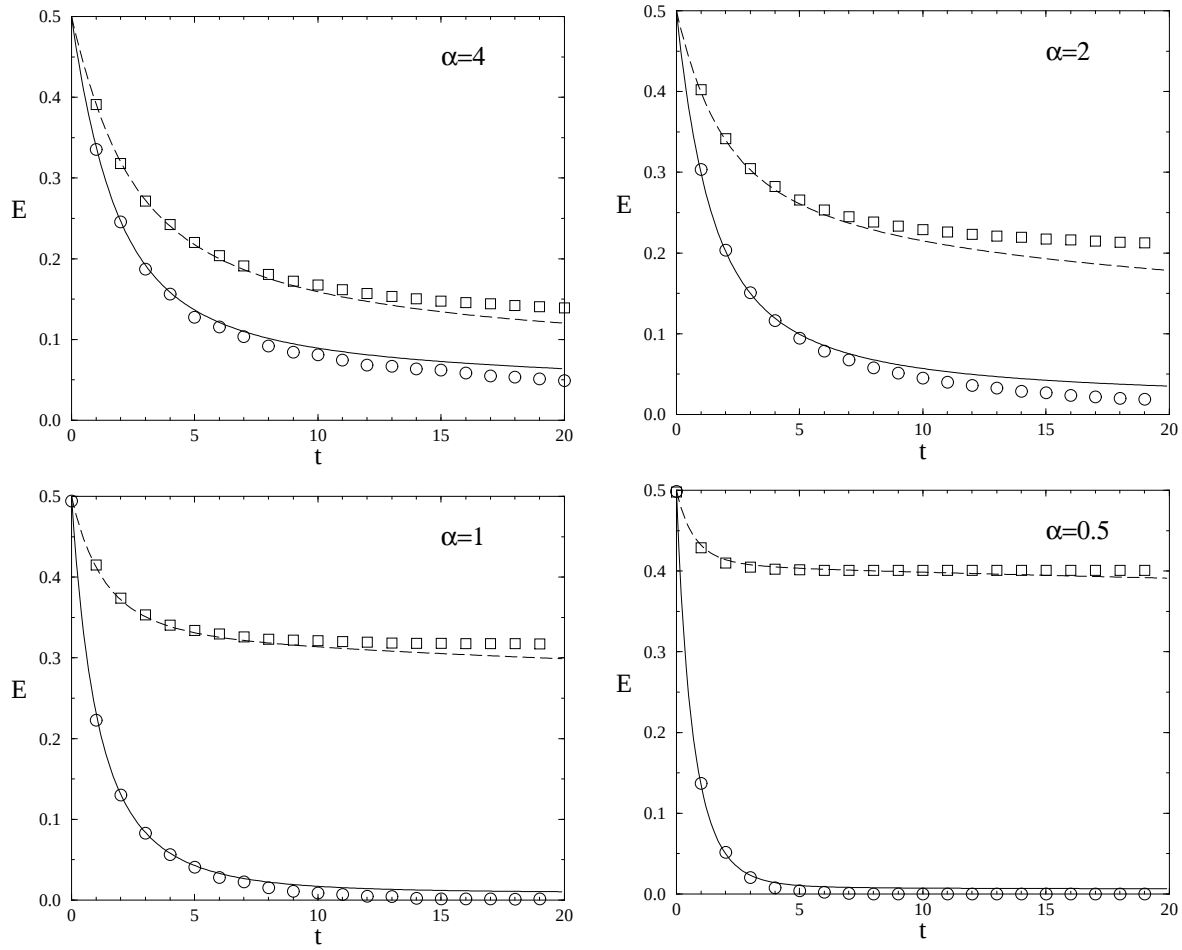

Figure 7: Comparison between the large α approximation of the theory and numerical simulations of on-line perceptron learning with $N = 10,000$ and $\eta = 1$. Markers: training errors $E_{\rm t}$ (circles) and generalisation errors $E_{\rm g}$ (squares); finite size effects in the simulation data are of the order of the marker size. Lines: theoretical predictions for training errors (solid) and generalisation errors (dashed) as functions of time, according to the approximated theory. Training set sizes: $\alpha = 4$ (upper left), $\alpha = 2$ (upper right), $\alpha = 1$ (lower left), and $\alpha = 0.5$ (lower right).

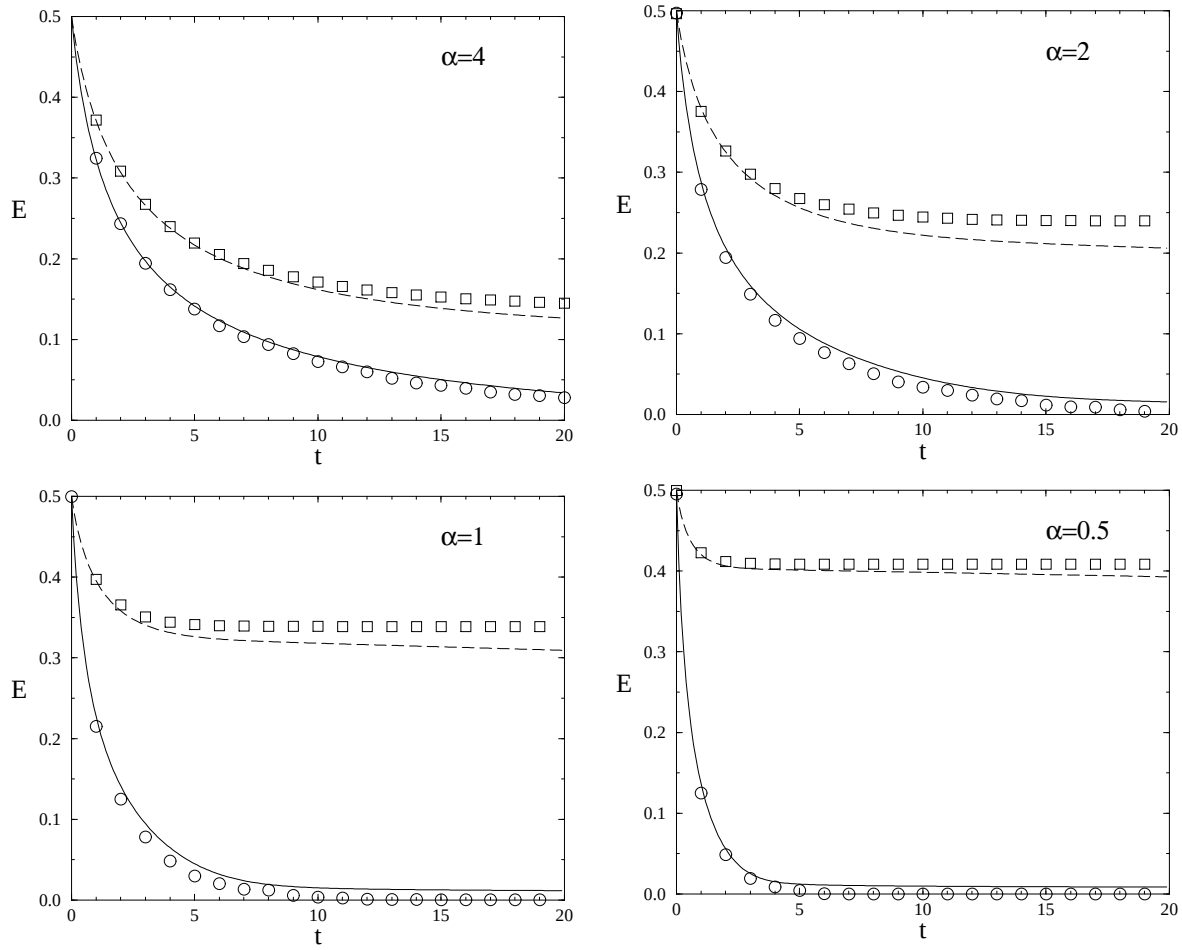

Figure 8: Comparison between the large α approximation of the theory and numerical simulations of on-line Adatron learning with $N = 10,000$ and $\eta = 2$. Markers: training errors $E_{\rm t}$ (circles) and generalisation errors $E_{\rm g}$ (squares); finite size effects in the simulation data are of the order of the marker size. Lines: theoretical predictions for training errors (solid) and generalisation errors (dashed) as functions of time, according to the approximated theory. Training set sizes: $\alpha = 4$ (upper left), $\alpha = 2$ (upper right), $\alpha = 1$ (lower left), and $\alpha = 0.5$ (lower right).

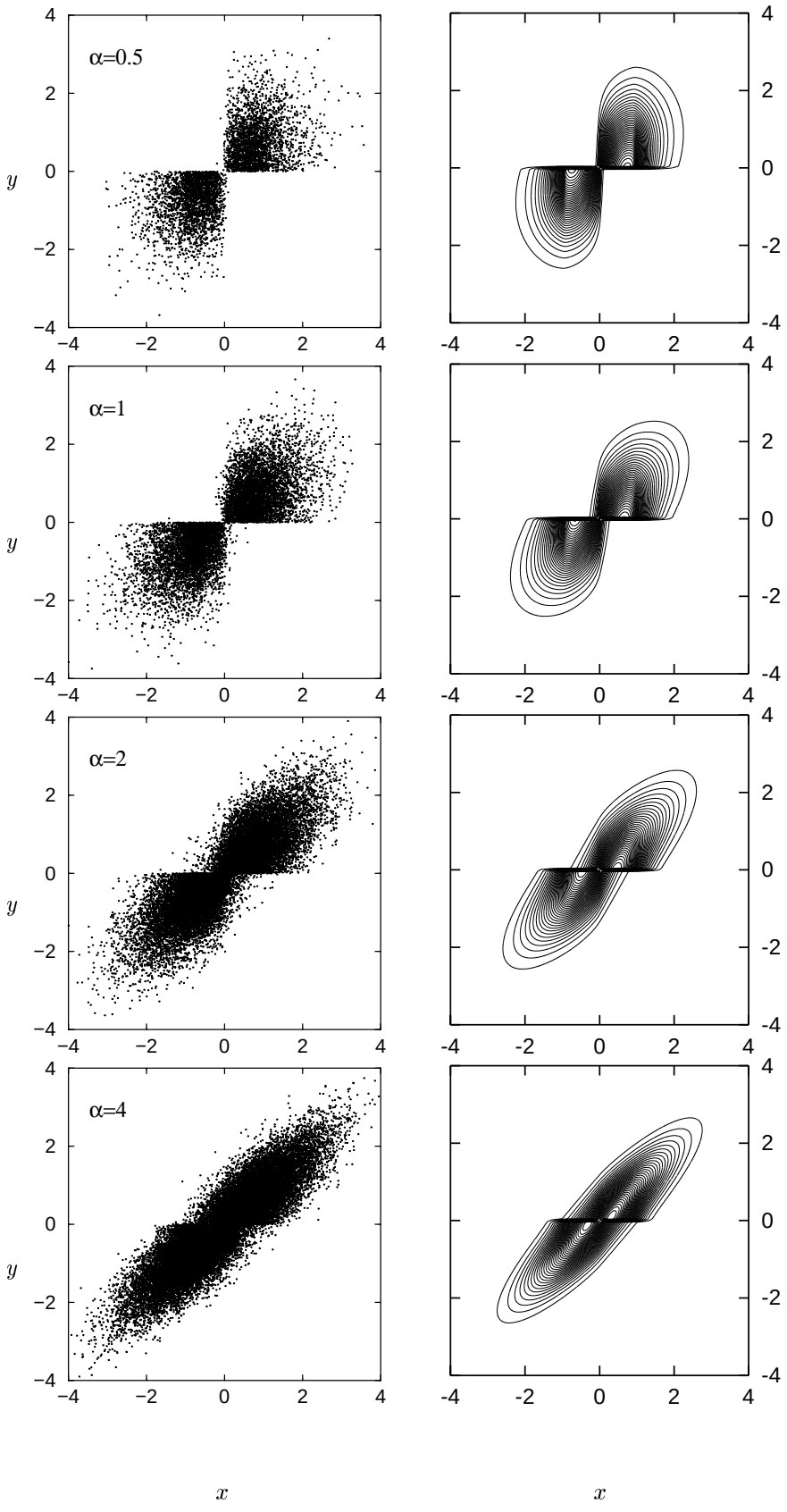

Figure 9: Comparison between the large α approximation of the theory and numerical simulations of on-line Perceptron learning, with $N = 10,000$ and $\eta = 1$. Scatter plots (left): observed student and teacher fields $(x, y) = (\boldsymbol{J}\cdot\boldsymbol{\xi}, \boldsymbol{B}\cdot\boldsymbol{\xi})$ as measured at time $t = 10$ during simulations, for the data in $\tilde{D}$, drawn in the (x, y) plane. Contour plots (right): corresponding predictions for the joint field distribution $P[x, y]$, according to the approximated theory. Training set sizes: $\alpha = 0.5, 1, 2, 4$ (from top to bottom).

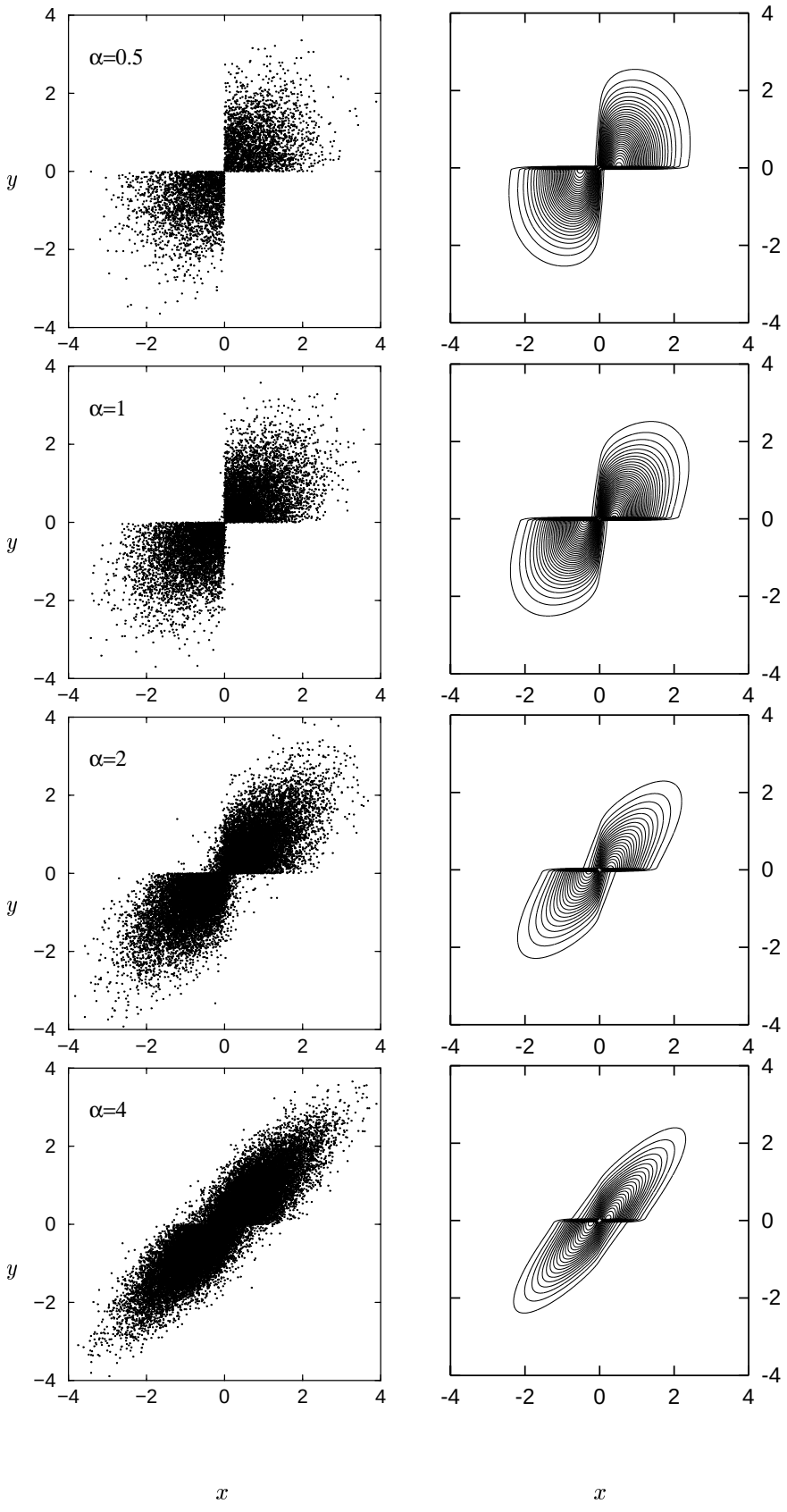

Figure 10: Comparison between the large α approximation of the theory and numerical simulations of on-line Adatron learning with $N=10,000$ and $\eta=2$. Scatter plots (left): observed student and teacher fields $(x,y)=(\boldsymbol{J}\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi})$ as measured at time $t=10$ during simulations, for the data in $\tilde{D}$, drawn in the (x,y) plane. Contour plots (right): corresponding predictions for the joint field distribution $P[x,y]$, according to the approximated theory. Training set sizes: $\alpha=0.5,1,2,4$ (from top to bottom).

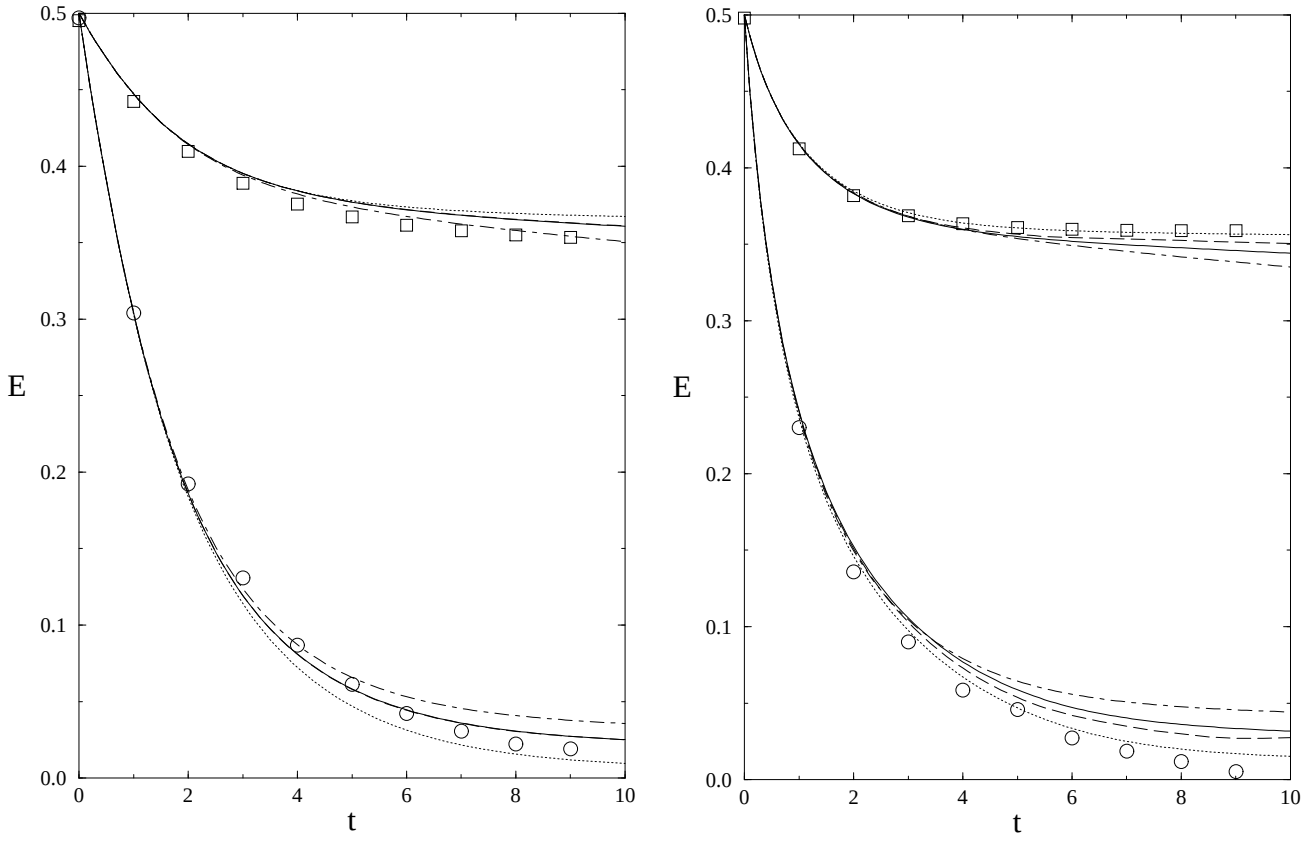

Figure 11: Comparison between the full numerical solution of our equations, as well as the three approximations of the theory, and the results of doing numerical simulations of on-line learning with $N = 10,000$ and $\alpha = 1$. Markers: training errors $E_{\rm t}$ (circles) and generalisation errors $E_{\rm g}$ (squares); finite size effects are of the order of the size of te markers. Lines: theoretical predictions for training errors (lower) and generalisation errors (upper) as functions of time, according to the theory. The different line types refer to: full equations (solid), annealed approximation (dashed), conditionally-Gaussian approximation (dashed-dotted) and large α approximation (dotted) (note: the dashed and solid curves fall virtually on top of one another). Left picture: Perceptron learning, with $\eta = \frac{1}{2}$. Right picture: AdaTron learning, with $\eta = \frac{3}{2}$.

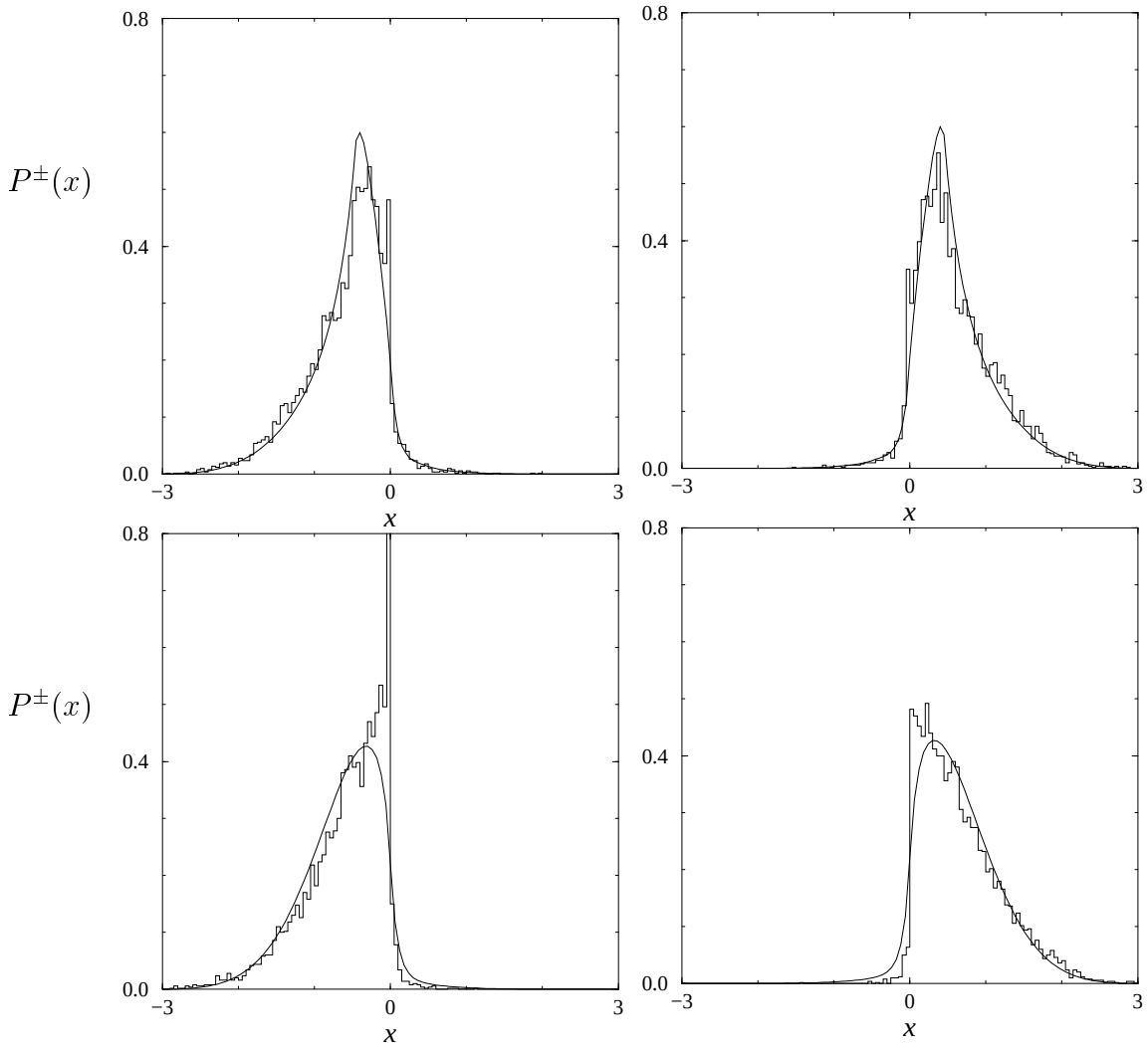

Figure 12: Comparison between the full numerical solution of our equations and the results of doing numerical simulations of on-line learning with $N = 10,000$ and $\alpha = 1$. Histograms: conditional student field distributions $P^{\pm}(x) = \int dy P[x, y]\theta[\pm y]$ as measured at time $t = 5$. Smooth curves: corresponding theoretical predictions. Upper pictures: Perceptron learning, with $\eta = \frac{1}{2}$ (left: $P^{-}(x)$, right: $P^{+}(x)$). Lower pictures: AdaTron learning, with $\eta = \frac{3}{2}$ (left: $P^{-}(x)$, right: $P^{+}(x)$).