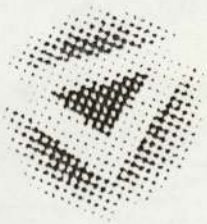
On-line learning in a changing environment

OTMAN BELMAHI

Master of Science by Research in Pattern Analysis and Neural Networks



THE UNIVERSITY OF ASTON IN BIRMINGHAM

December 1997

THE UNIVERSITY OF ASTON IN BIRMINGHAM

On-line learning in a changing environment

OTMAN BELMAHI

Master of Science by Research in Pattern Analysis and Neural Networks, 1997

Thesis Summary

On-line learning of non-stationary tasks by two-layer neural networks is studied within the framework of statistical mechanics. A fully connected network with K hidden units and fixed hidden-to-output weights (a soft committee machine) learns a non-stationary task represented by a network of similar architecture having M hidden nodes. The network is trained via gradient descent (standard back-propagation) on randomly drawn inputs and the corresponding outputs generated by the teacher network representing the task. This work employs a general framework for the dynamics of on-line learning obtained earlier for the fixed environment case. We describe a general task non-stationarity and investigate the learning process in these learning scenarios where on-line methods have been found to be most useful. The dynamics are first analysed for $K = M = 2$ which is the building block of the general case (any K and M).

The learning processes of stationary and non-stationary tasks are found to be qualitatively similar. However, for non-stationarities the transient stage of the dynamics becomes shorter and there is some residual error after convergence. These phases are investigated both numerically and analytically.

The insight gained from the non-stationary case leads to a new learning rule which seems to be more efficient than basic gradient descent in escaping the symmetric subspace related to the transient part of the dynamics. These effects are studied in arbitrary realisable scenarios ($K = M$).

Keywords: neural network, multi-layer perceptron, back-propagation, on-line learning, soft committee machine, non-stationary task

Acknowledgements

First I would like to thank David Saad and Magnus Rattray for their help answering my questions. I would also like to thank all the members of the Neural Computing Research Group for their kindness.

Contents

List of Figures

Chapter 1

Introduction

Neural networks approximate a system behaviour or a task by optimising parameters of a mathematical model [1]. As the optimisation process is highly time consuming and the performance and precision of a model cannot be determined in advance, the use of neural computing is confined to applications where efficient algorithmic solutions are impossible or impractical. Such applications are typically complex and poorly understood. Understanding speech, reading hand-written documents, and modelling and controlling non-linear systems are all areas where neural computing and other statistical techniques outperform algorithmic methods.

On-line learning is a popular method for training neural networks to identify the key features of the task to be learned. It extracts knowledge from each given example immediately rather than storing it for future use. This technique is particularly suitable for non-stationary systems i.e. systems whose parameters change in time, because data at a given time reflect a particular stage of the task rather than its general characteristics.

In contrast to most previous on-line learning studies which focus on stationary systems, this work concentrates on non-stationary tasks like [9]. To enable both numerical and analytical study, a statistical mechanics framework is employed and a generic non-stationarity of the system is presented, similarly to that used when investigating

learning abilities of stationary tasks [3].

1.1 Training neural networks

Multi-Layer Perceptrons (MLPs) are able to implement various input-output maps which are of importance to many classification and regression tasks. Two-layer architectures with N input units, one internal layer having an arbitrary unconstrained number of hidden units, and one output unit suffice to represent any scalar mappings of N-dimensional variables with arbitrary accuracy [2].

Internal parameters characterise a neural network of fixed architecture. Their choice determines specific maps $\zeta = f_W(\boldsymbol{\xi})$ from an N-dimensional input space $\boldsymbol{\xi}$ onto a scalar ζ (the index W is related to network internal parameters called weights). In order to bring the map f_W as close as possible to a desired map f_0, a process called training is used.

The process of learning from examples in layered neural networks is usually expressed as an optimisation problem, based on the minimisation of a training error computed over a training set composed of independent examples $(\boldsymbol{\xi}^\mu, \zeta^\mu)$. Network performance is measured by the generalisation error, which is the expected error on an unseen example. The two most common learning scenarios are batch and on-line.

In batch learning, training algorithms minimise the error calculated over the whole training set. There are a variety of efficient optimisation methods available, such as gradient descent or more sophisticated second order methods (e.g. Newton-Raphson or conjugate gradient)[1]. In on-line learning, single examples are presented sequentially and the training process adjusts the network parameters after the presentation of each example (e.g. using stochastic gradient descent)[1]. Here, the use of second order methods is complicated as the Hessian cannot be computed exactly and is only

approximated [1].

On-line methods are very often more efficient than batch in which costly computations and storage are required, especially for large data sets and input dimensions. Moreover, although it may be reasonable to store data generated by a fixed task and use it for training afterwards, it is less so for non-stationary mappings where examples reflect only a transient state of the process giving rise to data.

1.2 Motivations and objectives

Real world data are not all generated from stationary tasks and on-line learning is appropriate in such situations because it allows adaptation to learning mappings which change in time. However, most previous theoretical studies of on-line learning have indeed been concerned with stationary tasks. An extension of this work to non-stationary tasks is the subject of this project. In the present context, non-stationarity means that parameters of the mapping which generates training examples are being modified. This work employs the framework of statistical mechanics, which provides a compact description for the dynamics of on-line learning for stationary tasks [3]. The project focuses first on describing a possible non-stationarity which is as general as possible, and then on investigating the learning abilities of a two layer network trained on examples generated by a network of similar architecture. The various training phases are studied and compared to those found for a stationary task. As a byproduct of this study we find a method which improves learning abilities for stationary tasks by reducing the time required for escaping the transient stage of the dynamics which represents a significant part of the training process.

1.3 Thesis outline

The structure of the thesis is as follows:

- Chapter 2: After a presentation of the general framework, the derivation of the dynamics for learning stationary tasks is reproduced.
- Chapter 3: A general task non-stationarity is described. This is used to find the new dynamics, which can be integrated numerically and solved analytically in the neighbourhood of fixed points under certain assumptions.
- Chapter 4: A modified gradient descent rule is studied both numerically and analytically for stationary tasks to reduce the length of the symmetric plateau characterising the transient stage of the dynamics.
- Chapter 5: The different results obtained are summarised, and some possible extensions of the study are presented.

Chapter 2

The general framework

In this work on-line learning in MLPs is examined. To facilitate the learning process we are provided with a training set generated by the task to be learned. Since we are interested in a generic formulation of on-line learning, the outputs are supposed to be generated by a network of similar architecture but with possibly different complexity, in response to inputs drawn from a Gaussian distribution. The ability of a model network to learn the mapping provided by the network generating the data has been studied at length in [3] in the case of soft committee machine (two-layer networks with fixed hidden to output weights). Below we describe this model, which will be used later to study learning of non-stationary tasks.

2.1 Analytical description

We consider a learning scenario in which a model MLP is trying to learn a rule represented by another MLP on the basis of examples generated by the latter. The network which is being imitated represents the task and provides the training set, whereas the model network is being trained (i.e. its parameters are being modified) on the basis of these examples. It is useful to term them teacher and student networks respectively. In figure 2.1 these nets are shown graphically. Notice that they may have different number of hidden units $K \neq M$ (where K and M are the number of hidden nodes of the student

and teacher networks respectively). The network couplings from all hidden units to the output unit are set to 1. This special case can easily be extended to accommodate adaptive hidden to output weights [4] and preserves most properties of the general case.

Training examples are denoted $(\boldsymbol{\xi}^\mu, \zeta^\mu)$ with $\boldsymbol{\xi}^\mu$ describing N dimensional input vectors and ζ^μ the output of the teacher for the given input. Teacher nodes are associated with N dimensional vectors whose coordinates are the weights of the edges linking all the inputs to a hidden unit. The weight vector associated with teacher's node n is denoted $\mathbf{B}_n$ and the activation of this node is $y_n = \mathbf{B}_n \cdot \boldsymbol{\xi}$. Similarly, student nodes are associated with N dimensional weight vectors denoted $\mathbf{J}_i$; whose activation is $x_i = \mathbf{J}_i \cdot \boldsymbol{\xi}$. We use the index $i, j, k, \ldots$ to refer to the student and $n, m, \ldots$ for the teacher.

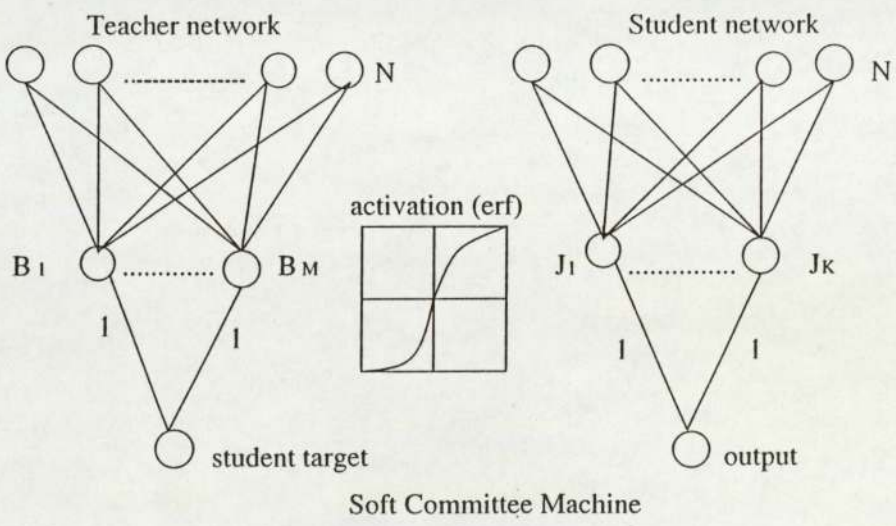


Figure 2.1: The task is represented by the teacher network which provides the targets of the student network. Each input value (N in total) is drawn from a normal Gaussian distribution for both networks. All the neurons have the same activation function which is the error function except the output node which is linear.

The output unit is linear and therefore the output value is $\sum_{i=1}^{K} g(x_i)$ for the student and $\sum_{n=1}^{M} g(y_n)$ for the teacher, where g represents the activation function of the hidden

nodes, which is taken to be an error function:

$$g(x) \equiv \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) = \sqrt{\frac{2}{\pi}} \int_0^x e^{-\frac{1}{2}t^2} \mathrm{d}t \ . \tag{2.1}$$

The error made by a student having weights $\mathbf{J}$ on a given input $\boldsymbol{\xi}$ is given by the quadratic deviation:

$$\epsilon(\mathbf{J}, \boldsymbol{\xi}) \equiv \frac{1}{2} \left[\sum_{i=1}^{K} g(x_i) - \sum_{n=1}^{M} g(y_n) \right]^2 \ . \tag{2.2}$$

As the components of each input vector $\boldsymbol{\xi}^\mu$ is drawn from a normal Gaussian distribution, the activations $\mathbf{x}$ and $\mathbf{y}$ fluctuate with the inputs. The distribution $P(\mathbf{x}, \mathbf{y})$ of the activation, where $\mathbf{x} = (x_1, ..., x_K)^T$ and $\mathbf{y} = (y_1, ..., y_M)^T$, is a multi-variate Gaussian:

$$P(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{(2\pi)^{M+K}|C|}} \exp\left\{ -\frac{1}{2} (\mathbf{x}, \mathbf{y})^T C^{-1} (\mathbf{x}, \mathbf{y}) \right\} \ , \tag{2.3}$$

the covariance matrix C is in terms of the overlaps among the weight vectors associated with the various hidden units as follows ($\langle . \rangle_\xi$ corresponds to an average over the inputs):

- $\langle x_i x_k \rangle_\xi = \mathbf{J}_i \ . \ \mathbf{J}_k \equiv Q_{ik}$ (between the i-th and k-th student units)
- $\langle x_i y_n \rangle_\xi = \mathbf{J}_i \ . \ \mathbf{B}_n \equiv R_{in}$ (between the i-th student unit and the n-th teacher unit)
- $\langle y_n y_m \rangle_\xi = \mathbf{B}_n \ . \ \mathbf{B}_m \equiv T_{nm}$ (between the n-th and m-th teacher units)

where

$$C = \begin{bmatrix} Q & R \\ R^T & T \end{bmatrix} \ . \tag{2.4}$$

The distribution is completely determined by Q_{ik}, R_{in}, and T_{nm}. These elements are called order parameters and are sufficient to represent key features of the learning process. The parameters T_{nm} are characteristic of the task to be learned and we mostly consider an isotropic teacher for which $\mathbf{B}_n \ . \ \mathbf{B}_m = \delta_{nm}$. The overlaps Q_{ik} among student hidden units and R_{in} between student and teacher hidden units are determined by the student weights $\mathbf{J}$ and evolve during training.

A gradient descent rule for the update of student weights is used:

$$\mathbf{J}^{\mu+1} = \mathbf{J}^{\mu} - \frac{\eta}{N} \nabla_{\mathbf{J}} \, \epsilon \, (\mathbf{J}^{\mu}, \boldsymbol{\xi}^{\mu}) \, , \tag{2.5}$$

where the learning rate η is scaled with the input size N in order to take account for the fluctuations of the variables. The role of η is to determine the speed of the training process. By calculating $\nabla_{\mathbf{J}}$ explicitly for this model, the rule becomes:

$$\mathbf{J}_i^{\mu+1} = \mathbf{J}_i^{\mu} + \frac{\eta}{N} \, \delta_i^{\mu} \, \boldsymbol{\xi}^{\mu} \, , \tag{2.6}$$

where

$$\delta_i^{\mu} \equiv g'(x_i^{\mu}) \left[\sum_{n=1}^{M} g(y_n^{\mu}) - \sum_{j=1}^{K} g(x_j^{\mu}) \right] \, , \tag{2.7}$$

is defined in terms of both the activation function and its derivatives g'. The time evolution of $R_{in} = \mathbf{J}_i \, . \, \mathbf{B}_n$ is then given by:

$$R_{in}^{\mu+1} - R_{in}^{\mu} = \frac{\eta}{N} \, \delta_i^{\mu} \, y_n^{\mu} \, , \tag{2.8}$$

similarly for $Q_{ik} = \mathbf{J}_i \, . \, \mathbf{J}_k$, it is:

$$Q_{ik}^{\mu+1} - Q_{ik}^{\mu} = \frac{\eta}{N} \, (\delta_i^{\mu} \, x_k^{\mu} + \delta_k^{\mu} \, x_i^{\mu}) + \frac{\eta^2}{N^2} \, \delta_i^{\mu} \, \delta_k^{\mu} \, \boldsymbol{\xi}^{\mu} . \, \boldsymbol{\xi}^{\mu} \, . \tag{2.9}$$

These equations, which are valid for any stationary task (i.e. $\mathbf{B}_n$ does not depend on time) are discrete. The terms on the right-hand side of the dynamical equations are fluctuating with the inputs. We are interested, however, in the mean behaviour of the network and averages of the observed quantities are therefore calculated. These averages should be computed over the inputs $\boldsymbol{\xi}$, but it is entirely equivalent to calculate them over the probability distribution of the activations as all relevant quantities depend on the activations. Averaging the training error over all examples constitutes the generalisation error:

$$\epsilon_g(\mathbf{J}) \equiv \quad \langle \epsilon(\mathbf{J}, \boldsymbol{\xi}) \rangle_{\{\xi\}} \, , \tag{2.10}$$

and its expression is in terms of the order parameters only [3]:

$$\epsilon_g(\mathbf{J}) = \frac{1}{\pi}\left\{\sum_{ik}\arcsin\frac{Q_{ik}}{\sqrt{1+Q_{ii}}\sqrt{1+Q_{kk}}}+\sum_{nm}\arcsin\frac{T_{nm}}{\sqrt{1+T_{nn}}\sqrt{1+T_{mm}}}\right.$$
$$\left.-2\sum_{in}\arcsin\frac{R_{in}}{\sqrt{1+Q_{ii}}\sqrt{1+T_{nn}}}\right\}, \tag{2.11}$$

where $1 \leq i,k \leq K$ sum over the student hidden units, and $1 \leq n,m \leq M$ sum over the teacher hidden units. It is also more convenient to work with continuous rather than discrete dynamics. The time used is $t = \frac{\mu}{N}$ and the equations of R can then be written:

$$\frac{\Delta R_{in}}{\Delta t} \equiv \frac{R_{in}^{\mu+1} - R_{in}^{\mu}}{1/N} = \eta\ \delta_i^{\mu} y_n^{\mu}\ . \tag{2.12}$$

By considering N very large, Δt becomes very small and the previous equation can then be written in a differential form:

$$\frac{dR_{in}}{dt} = \eta\ \langle \delta_i y_n \rangle\ . \tag{2.13}$$

The equations for Q's are derived in the same way by considering equation 2.9:

$$\frac{\Delta Q_{ik}}{\Delta t} \equiv \frac{Q_{ik}^{\mu+1} - Q_{ik}^{\mu}}{1/N} = \eta(\delta_i^{\mu} x_k^{\mu} + \delta_k^{\mu} x_i^{\mu}) + \frac{\eta^2}{N}\langle \delta_k^{\mu}\delta_i^{\mu}\boldsymbol{\xi}^{\mu}.\ \boldsymbol{\xi}^{\mu}\rangle_{\xi}\ . \tag{2.14}$$

Each input neuron is drawn from a normal Gaussian (zero mean and unit variance), therefore we can assert that:

$$\langle \boldsymbol{\xi}^{\mu}.\ \boldsymbol{\xi}^{\mu}\rangle_{\xi} \equiv N\ , \tag{2.15}$$

and :

$$\frac{\Delta Q_{ik}}{\Delta t} \equiv \eta(\delta_i^{\mu} x_k^{\mu} + \delta_k^{\mu} x_i^{\mu}) + \eta^2 \delta_k^{\mu}\delta_i^{\mu}\ . \tag{2.16}$$

The large N hypothesis is called thermodynamic limit and allows one to neglect the variance of the fluctuations which are $O(\frac{1}{\sqrt{N}})$, so that the average is sufficient to represent the dynamics. The equations found are [3]:

$$\begin{aligned}\frac{dR_{in}}{dt} &= \eta\ \langle \delta_i\ y_n\rangle \\ \frac{dQ_{ik}}{dt} &= \eta\ \langle \delta_i\ x_k\ +\ \delta_k\ x_i\rangle +\ \eta^2\ \langle \delta_i\ \delta_k\rangle\ ,\end{aligned} \tag{2.17}$$

The averages in the system above require the evaluation of two types of multivariate Gaussian integrals:

- $I_3 \equiv \langle g'(u)\ v\ g(w)\rangle$, for terms proportional to η, where the argument u of g' is one of the components of $\mathbf{x}$, while both v and w can be components of either $\mathbf{x}$ or $\mathbf{y}$.

- $I_4 \equiv \langle g'(u)\ g'(v)\ g(w)\ g(z)\rangle$, for terms proportional to η^2, where u and v are components of $\mathbf{x}$ while w and z can be components of either $\mathbf{x}$ or $\mathbf{y}$.

These averages have been calculated in closed form, and only involve the overlaps [3] (see Appendix A).

2.2 Numerical solutions

To obtain a full description of the learning dynamics we solve the system of equations (2.17) numerically (using Runge and Kutta technique). Initial conditions are selected to reflect our lack of knowledge about the task, i.e. to represent randomly chosen student vectors. Throughout this work we will use randomly drawn values from a uniform distribution for both initial student-teacher and student-student overlaps.

The equations are solved numerically for the case of a two hidden nodes network learning an isotropic teacher ($T_{nm} = \delta_{nm}$) with the same architecture ($K = M = 2$). The evolution of the student-student overlaps (Q_{ij}) and the student-teacher overlaps (R_{in}) are shown in figure 2.2. These curves show two plateaus, one characterised by $Q_{ij} \approx Q\ \delta_{ij} + (1 - \delta_{ij})\ C$, $R_{in} \approx R\ \forall i, j, n$, and another one where $Q_{ij} \approx \delta_{ij}$, $R_{in} \approx \delta_{in}\ \forall i, j, n$, when the indices have been ordered appropriately. The first plateau for Q's and R's corresponds to a transient of the dynamics and the generalisation error is much larger than zero (see figure 2.3). This stage is called the symmetric phase and is characterised by a lack of differentiation between different student nodes. The second phase corresponds to exponential convergence of the order

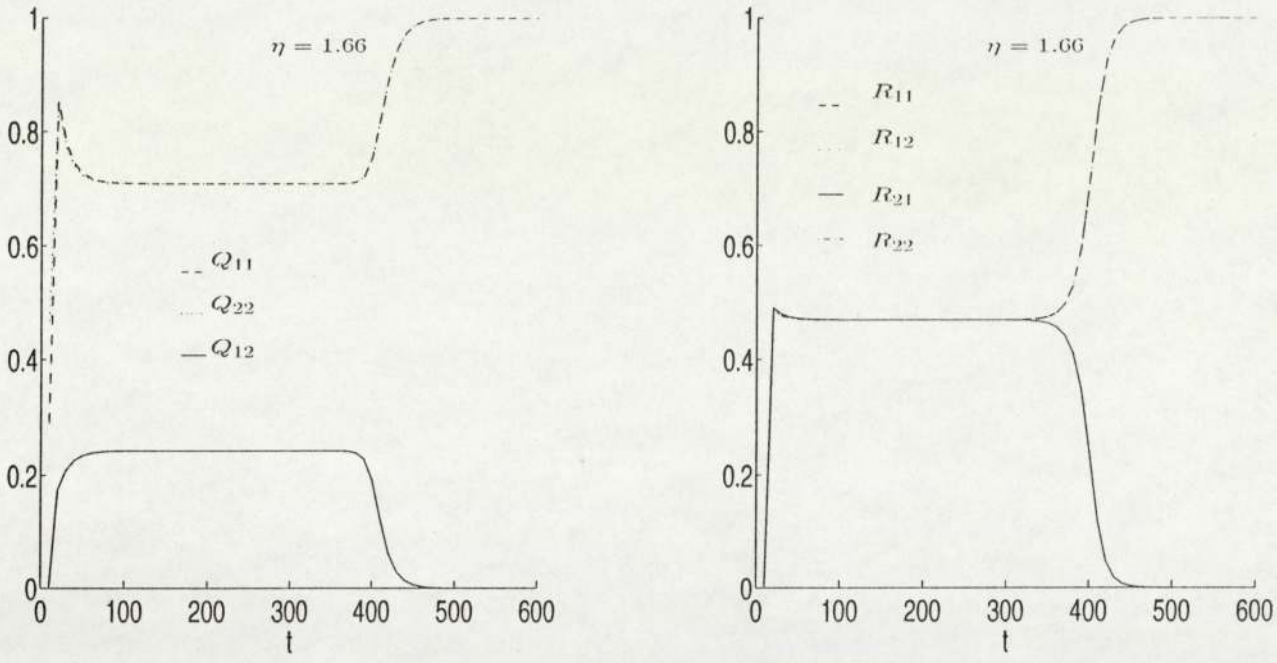


Figure 2.2: Evolution of the order parameters Q_{ij}, R_{in} for an architecture $K = M = 2$ (η=1.66) and an isotropic teacher. Two stages can be distinguished and correspond to the symmetric and convergence phase of the learning. Initial conditions are drawn from a uniform distribution.

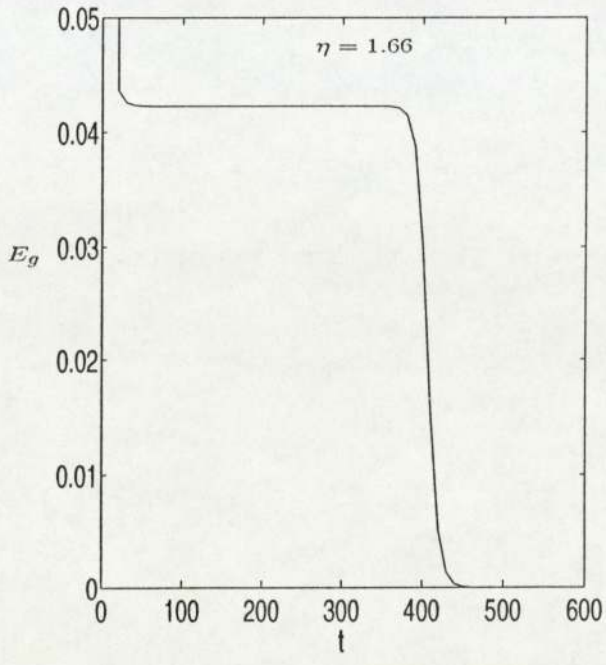


Figure 2.3: Evolution of the generalisation error for an architecture $K = M = 2$ (η=1.66) and an isotropic teacher (T_{nm}=δ_{nm}). Two stages are observed and correspond to those of the order parameters.

parameters to their optimal values. The generalisation error also exhibits an exponential decay towards zero (see figure 2.3) . This is termed the convergence phase and is characterised by the specialisation of each student node to a specific teacher node.

Student vectors have different behaviours in each phase. They point towards the same direction within the teacher subspace during the symmetric phase and each one starts to point in the direction of a different teacher vector at the beginning of the convergence phase. Finally the student vectors become aligned with those of the teacher [3].

Specialisation is then a characteristic of the convergence phase only, although the system spends a long time in the symmetric subspace before escaping. For this reason, a number of studies have suggested modified training algorithms for getting rid off the symmetric phase, or at least reducing it significantly [5,6]. In chapter 4 we will suggest another modification, which is also shown to reduce this phase considerably.

2.3 Summary

A statistical mechanics framework is used to describe the learning process. For a realisable stationary task (same number of hidden nodes in student and teacher networks where teacher vectors do not depend on time) we find two distinct stages called the symmetric and convergence phases respectively. Specialisation is the consequence of the convergence phase only. We now wish to apply this framework to a changing environment and this is the subject of the next chapter.

Chapter 3

On-line learning in a changing environment

In order to derive the dynamics for learning non-stationary tasks, we need first a framework and then a complete description of the non-stationarity. A framework similar to that used for stationary systems is employed in conjunction with a generic description of task non-stationarity, in order to have as general result as possible.

Below we derive the dynamics for a certain non-stationarity and compare the numerical and analytical results obtained to those corresponding to learning stationary tasks [3].

3.1 Modelling changes in the environment

There are many possibilities for modelling temporal changes of the task to be learned. A particular example is the smooth rotation which we will focus on since it is amenable to theoretical analysis and may have some similar features to non-stationary tasks for which on-line learning is useful. Generally, teacher vectors can have all sorts of cross-correlations; here we will restrict our study to the case of orthonormal vectors whose non-stationarity is characterised by a single parameter. The fact that these high dimensional vectors are chosen randomly motivates the orthonormality assumption

which simplifies the analytical work. Furthermore, we will restrict ourselves to scenarios where both student and teacher have the same number of hidden nodes ($K = M$). Two restrictions on the task non-stationarity are then made:

- smooth changes.
- teacher vectors remain orthonormal.

In order to study the dynamics, changes affecting teacher vectors should be explicit because of their influence on the equations for the dynamics. It is convenient to consider a linear non-stationarity of the form $\Psi : \mathcal{V} \rightarrow \mathcal{V}$, where $\mathcal{V}$ is a vectorial space of finite dimension and the mapping is continuous. Teacher vectors remain orthonormal, so Ψ transforms from one orthonormal basis to another i.e. Ψ is an orthogonal transformation. The determinant of such a transformation is either $+1$ or -1, which makes the distinction between:

- symmetries, with a determinant equal to -1,
- rotations, with a determinant equal to $+1$.

Symmetry is inherently a discrete transformation and therefore will not be considered in this study. However, rotation is smooth as it can be defined in terms of infinitesimal processes. Rotating a vector with an angle ω is equivalent to rotating it n times with an angle $\frac{\omega}{n}$ each and smoothness results from this property.

Teacher vectors are then rotated slowly in the N dimensional space. A rotation of a set of vectors is not unique and we consider one type of rotation first. The initial teacher space is denoted $\mathcal{T}=vect(\mathbf{B}_1,...,\mathbf{B}_K)$, the whole space is represented by $\mathcal{V}=\mathcal{T} \oplus \mathcal{T}^{\perp}$. The rotation is N dimensional but it can occur either in a subspace (for instance $\mathcal{T}$) or in the full space of dimension N. The interest of this study is mainly in identifying student's learning abilities when teacher vectors are changing and we will therefore focus on a simple case where rotations are in the initial teacher's space $\mathcal{T}$. Moreover, according to [3] (for stationary teacher vectors), the relevant dynamics are

mainly in the space $\mathcal{T}$ while dynamics in $\mathcal{T}^{\perp}$ are an artifact of the stochasticity of the learning process, which is influenced by the choice of training parameters (e.g. η).

Rotations are then restricted to the subspace spanned by the initial position of teacher vectors whose coordinates can be described in terms of a K dimensional vector:

$$\mathbf{B}_n = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \end{pmatrix} \quad n\,element$$

Teacher vectors are being rotated and as mentioned before, it is better to find a general transformation rather than a particular one. One can look at the reduction of a rotation in an orthonormal basis.

Each rotation $\mathcal{A}$ can be represented by a matrix:

$$\tilde{\mathbf{A}} = \begin{pmatrix} 1 & 0 & & \cdots & & 0 \\ 0 & \ddots & & & & \\ & & 1 & & 0 & \\ \vdots & & & \boxed{A_1} & & \vdots \\ & & 0 & & \ddots & 0 \\ 0 & & \cdots & & 0 & \boxed{A_p} \end{pmatrix} \tag{3.1}$$

in an orthonormal basis. The elements A_1 to A_p represent rotations in fixed planes and therefore their expression is given by:

$$\mathbf{A_i} = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix} \tag{3.2}$$

for a very small rotation frequency (i.e. small θ_i), the matrix becomes:

$$\mathbf{A_i} = \begin{pmatrix} 1 & -\theta_i \\ \theta_i & 1 \end{pmatrix} \tag{3.3}$$

The proof of the above result is given in appendix B. This reduction shows that a general rotation can be written as a set of 2×2 rotations. The building block of a general rotation is therefore the $K = M = 2$ case studied below.

3.2 Learning dynamics and the $K = M = 2$ case

The framework is now restricted to the case $K = M = 2$ in which a 2 hidden nodes student network is trying to learn a 2 hidden nodes non-stationary teacher where the vectors rotate but remain orthonormal. By assuming a very slow rotation, the non-stationarity is described by a transformation:

$$\mathcal{A} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{I} + \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix} \tag{3.4}$$

where θ represents the frequency of the rotation. Its order has not been defined yet and is of importance for the equations derived later.

The learning rule (equation 2.6) uses a scaled learning rate $\frac{\eta}{N}$ to control the fluctuations. Similarly, the rotation should appear with a $\frac{1}{N}$ scaling to preserve the smooth evolution of the system, and therefore we denote $\theta = \frac{\omega}{N}$. It is important to notice that the rotation is not a fluctuating quantity but to maintain valid dynamics, the learning rate and the frequency should have the same order in N. These orders enable us to derive smooth continuous differential equations for the order parameters. The matrix

$\mathcal{A}$ becomes:

$$\mathcal{A} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{I} + \underbrace{\begin{pmatrix} 0 & -\frac{\omega}{N} \\ \frac{\omega}{N} & 0 \end{pmatrix}}_{A^N} . \tag{3.5}$$

The expressions for the new dynamics can be derived under the hypothesis that teacher vectors are continuously modified by $\mathcal{A}$. The correlation between student and teacher vectors is given by: $R_{in}^{\mu} = \mathbf{J}_i^{\mu} \cdot \mathbf{B}_n^{\mu}$ for the example μ. Therefore it is possible to write:

$$\begin{aligned} R_{in}^{\mu+1} &= (\mathbf{J}_i^{\mu} + \frac{\eta}{N}\, \delta_i^{\mu} \boldsymbol{\xi}^{\mu}) \cdot \mathbf{B}_n^{\mu+1} \\ &= \mathbf{J}_i^{\mu} \cdot \mathbf{B}_n^{\mu} + \mathbf{J}_i^{\mu} \cdot \sum_{j=1}^{2} A_{n,j}^{N}\, \mathbf{B}_j^{\mu} + \frac{\eta}{N}\, \delta_i^{\mu} \boldsymbol{\xi}^{\mu} \cdot \mathbf{B}_n^{\mu} + \frac{\eta}{N} \delta_i^{\mu} \boldsymbol{\xi}^{\mu} \cdot \sum_{j=1}^{2} A_{n,j}^{N}\, \mathbf{B}_j^{\mu} , \end{aligned}$$

where $A_{n,j}^{N}$ indicates the element (n, j) of the matrix A^N. By considering the difference in R before and after introducing example μ,

$$\frac{R_{in}^{\mu+1} - R_{in}^{\mu}}{1/N} = \eta\, \delta_i^{\mu} y_n^{\mu} + \mathbf{J}_i^{\mu} \cdot \underbrace{\sum_{j=1}^{2} N\, A_{n,j}^{N}\, \mathbf{B}_j^{\mu}}_{O(1)} + \eta\, \underbrace{\delta_i^{\mu} \boldsymbol{\xi}^{\mu} \cdot \sum_{j=1}^{2} A_{n,j}^{N}\, \mathbf{B}_j^{\mu}}_{O(1/N)} . \tag{3.6}$$

By considering $t = \frac{\mu}{N}$ as the time, the expression can be written as:

$$\frac{\Delta R_{in}}{\Delta t} = \eta\, \delta_i^{\mu} y_n^{\mu} + \mathbf{J}_i^{\mu} \cdot \sum_{j=1}^{2} N\, A_{n,j}^{N}\, \mathbf{B}_j^{\mu} + O\left(\frac{1}{N}\right) . \tag{3.7}$$

The thermodynamic limit is applied ($N \to \infty$) which means that the number of inputs becomes very large while the number of hidden nodes is fixed. Terms of $O(\frac{1}{N})$ are then neglected compared to terms of $O(1)$, and we obtain a set of differential equations:

$$\frac{dR_{in}}{dt} = \eta\, \langle \delta_i y_n \rangle_{x,y} + \mathbf{J}_i \cdot \sum_{j=1}^{2} N\, A_{n,j}^{N}\, \mathbf{B}_j^{\mu} . \tag{3.8}$$

A similar treatment is carried out for the correlations between student vectors (Q_{ij}) showing that the result for a stationary environment in equation (2.17) remains the same. We denote P^{ω} the relevant order parameter expression for a rotation with angle

ω. It is possible to write the dynamical equations for the $K = M = 2$ case:

$$\begin{aligned}
\frac{dR_{11}^{\omega}}{dt} &= \frac{dR_{11}^{0}}{dt} - \omega\, R_{12}^{\omega} \\
\frac{dR_{21}^{\omega}}{dt} &= \frac{dR_{21}^{0}}{dt} - \omega\, R_{22}^{\omega} \\
\frac{dR_{12}^{\omega}}{dt} &= \frac{dR_{12}^{0}}{dt} + \omega\, R_{11}^{\omega} \\
\frac{dR_{22}^{\omega}}{dt} &= \frac{dR_{22}^{0}}{dt} + \omega\, R_{21}^{\omega} \\
\frac{dQ_{11}^{\omega}}{dt} &= \frac{dQ_{11}^{0}}{dt}\,,\ \frac{dQ_{12}^{\omega}}{dt} = \frac{dQ_{12}^{0}}{dt}\,,\ \frac{dQ_{12}^{\omega}}{dt} = \frac{dQ_{22}^{0}}{dt}\,.
\end{aligned} \tag{3.9}$$

The expressions of $\frac{dP^0}{dt}$ are given in equation 2.17 for the stationary case. The initial conditions used for solving the equations numerically are similar to those used in section 2.2. Learning abilities are then studied in this context. The behaviour in the case of $\omega = 10^{-5}$ is qualitatively similar to the one observed in section 2.2. There are still two main stages of learning, the symmetric and convergence phases as observed in figure 3.1.

Note that the symmetric plateau has a length around 100 in comparison to 400 in the fixed environment case (figure 2.2). The plateau is then much shorter when the task is non stationary which is shown clearly in figure 3.2 where the generalisation error is plotted for $\omega = 10^{-5}$ and $\omega = 0$.

For larger values of ω, the task becomes unlearnable as shown in figure 3.3. The behaviour is then dependent on the rotation rate ω and to study the effect of non-stationarity on the symmetric and convergence phases we resort to an analytical study with some similarities to [3], based on expansions around the symmetric and asymptotic fixed point and then on considering the linear systems obtained.

3.2.1 Symmetric phase

When the environment is stationary [3], all student vectors in the symmetric phase are pointing towards the same direction and show no specialisation, the generalisation error is constant but much larger than zero. This phase is characterised by:

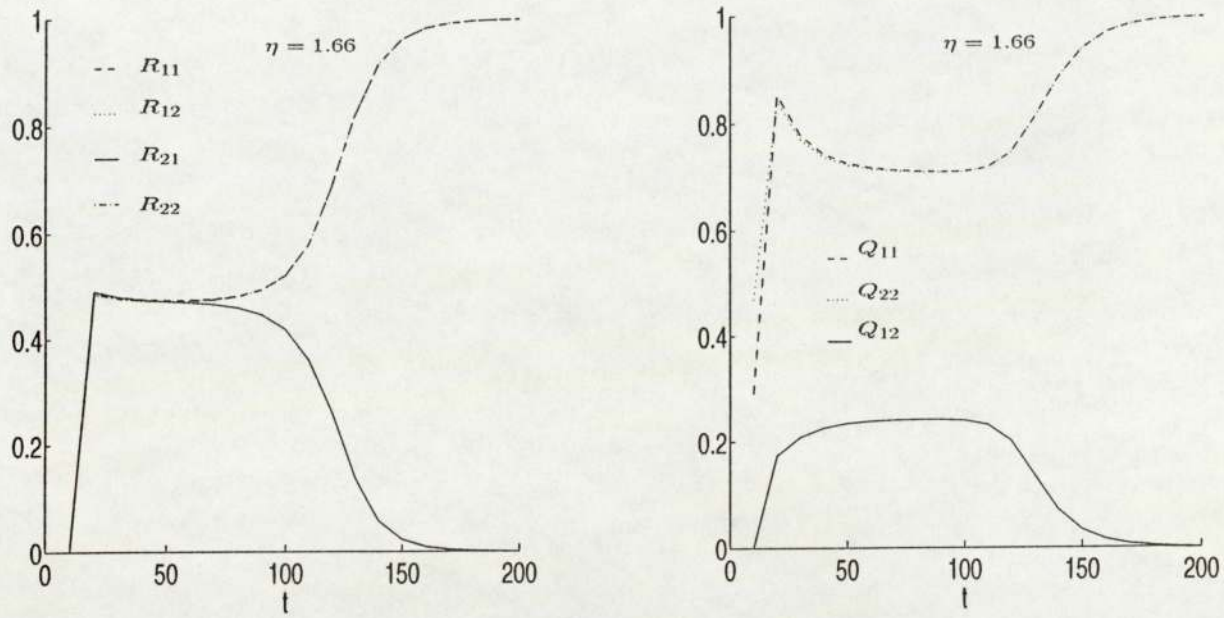


Figure 3.1: Dynamical evolutions of the order parameters for $\omega = 10^{-5}$, $\eta = 1.66$, $K = M = 2$. The dynamics are qualitatively similar to those observed in the fixed environment case.

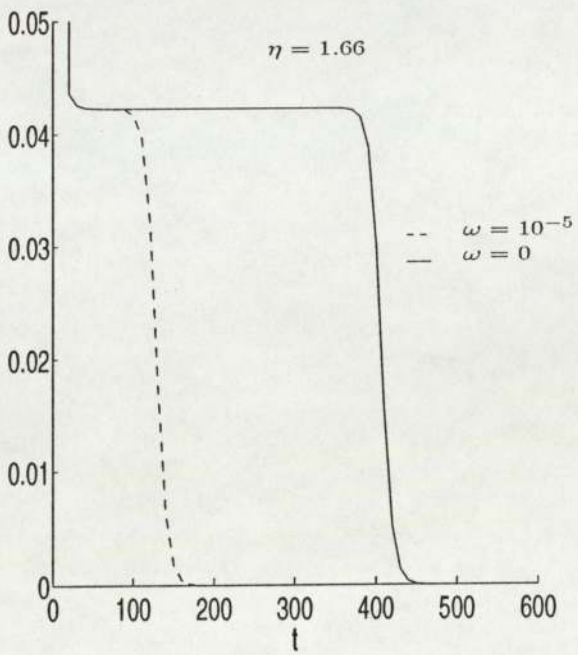


Figure 3.2: Generalisation error of the $K = M = 2$ case for the fixed environment ($\omega = 0$) and rotating tasks ($\omega = 10^{-5}$). These two curves are drawn for the same learning rates, $\eta = 1.66$. The symmetric plateau is shorter when the task is non stationary.

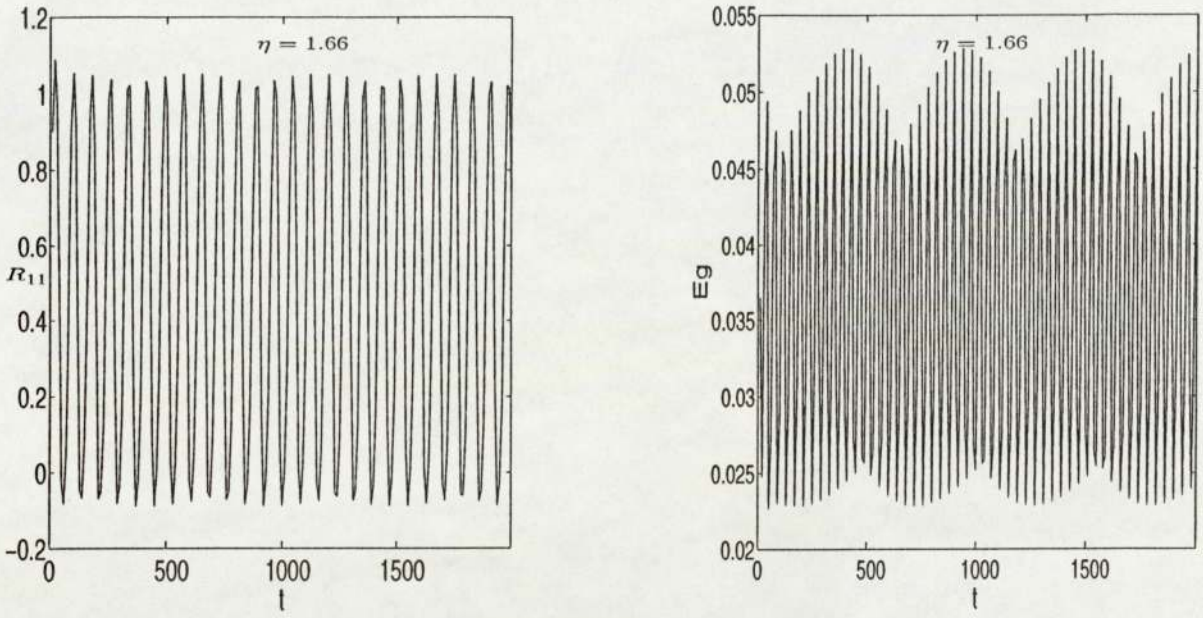


Figure 3.3: Dynamical evolution of R_{11} and of ϵ_g for $\omega = 0.1$, $K = M = 2$. The task is unlearnable as the generalisation error is oscillating and remain high.

- An unstable symmetric fixed point (in terms of the order parameters) of the dynamics, which corresponds to the values at the plateaus observed in the evolution of the overlaps between teacher-student and student-student nodes
- A corresponding plateau in the generalisation error
- The symmetric subspace is escaped by the student-teacher overlaps (R_{in}) first, and then by the student-student correlation.

As a first step we should find the new fixed point which is not at the same position as for stationary tasks (figures 2.2 and 3.1). As the frequency of the rotation is supposed to be very small, the new fixed point will be close to the one found in [3], and can be obtained by expanding around the stationary fixed point.

Initially there are four R overlaps:

- R_{11} and R_{21}, representing the projection of $\mathbf{J}_1$ and $\mathbf{J}_2$ on $\mathbf{B}_1$
- R_{12} and R_{22}, representing the projection of $\mathbf{J}_1$ and $\mathbf{J}_2$ on $\mathbf{B}_2$

Students vectors lie in the whole space of dimension N. As the teacher's orthogonal space is related mainly to the stochasticity of the learning process, its influence can be neglected if we assume a small learning rate which reduces the effect of fluctuations. Therefore we can assume that student vectors lie in the teacher space and the four teacher-student overlaps can then be reduced to two using the relations:

- $R_{12} = \sqrt{Q_{11} - R_{11}^2}$
- $R_{22} = \sqrt{Q_{22} - R_{21}^2}$

The two variables (R_{11} and R_{21}) then describe the behaviour of all teacher-student correlations. To simplify the notations, we define $R = R_{11}$ and $S = R_{21}$. To reduce the number of parameters further, two different categories can be distinguished in the student-student correlations: Q_{ii} which is the squared length of the i-th student vector and Q_{ij} the inner product between $\mathbf{J}_i$ and $\mathbf{J}_j$ ($i \neq j$). It is therefore possible to write down: $Q_{ij} = Q\ \delta_{ij} + (1 - \delta_{ij})\ C$. There are then only two variables representing the student correlations (Q and C) instead of the original three (Q_{11},Q_{12} and Q_{22}). The reduction of the number of variables assumes a small learning rate value. This is motivated by the fact that the symmetric phase is more emphasised for small η. This assumption enables us to neglect terms proportional to η^2 in the dynamics presented in equation (2.17) and to exploit the simple relations between R and Q. The terms proportional to η will then be considered for finding the symmetric fixed point. The symmetric fixed point is defined when the system 3.10 is zero.

$$
\begin{aligned}
\frac{dR}{dt} &= \frac{2\eta}{\pi(1+Q)}\left(\frac{1+Q-R^2}{\sqrt{2(1+Q)-R^2}} - \frac{R\sqrt{Q-R^2}}{\sqrt{2+Q+R^2}} - \frac{R}{\sqrt{1+2Q}} - \frac{S(1+Q)-RC}{\sqrt{(1+Q)^2-C^2}}\right) - \omega\ S, \\
\frac{dS}{dt} &= \frac{2\eta}{\pi(1+Q)}\left(\frac{1+Q-S^2}{\sqrt{2(1+Q)-S^2}} - \frac{S\sqrt{Q-S^2}}{\sqrt{2+Q+S^2}} - \frac{S}{\sqrt{1+2Q}} - \frac{R(1+Q)-SC}{\sqrt{(1+Q)^2-C^2}}\right) - \omega\ R, \\
\frac{dQ}{dt} &= \frac{4\eta}{\pi(1+Q)}\left(\frac{R}{\sqrt{2(1+Q)-R^2}} + \frac{\sqrt{Q-R^2}}{\sqrt{2+Q+R^2}} - \frac{Q}{\sqrt{1+2Q}} - \frac{C}{\sqrt{(1+Q)^2-C^2}}\right), \qquad (3.10) \\
\frac{dC}{dt} &= \frac{2\eta}{\pi(1+Q)}\left(\frac{S(1+Q)-CR}{\sqrt{2(1+Q)-R^2}} + \frac{\sqrt{Q-S^2}(1+Q)-C\sqrt{Q-R^2}}{\sqrt{2+Q+R^2}} + \frac{R(1+Q)-CS}{\sqrt{2(1+Q)-S^2}}\right. \\
&+ \left.\frac{(1+Q)\sqrt{Q-R^2}-C\sqrt{Q-S^2}}{\sqrt{2+Q+S^2}} - \frac{2C}{\sqrt{1+2Q}} - 2\frac{Q^2-C^2+Q}{\sqrt{(1+Q)^2-C^2}}\right).
\end{aligned}
$$

This is a non linear system in terms of the four variables R, S, Q and C. Finding an analytical solution directly is a complex task. However, as the non stationarity is

supposed to be small it may be assumed that the new fixed point is close to the one observed for a fixed environment which is given analytically by [3]:

$$\begin{aligned} Q^{0^*} &= C^{0^*} = \frac{1}{3} , \\ R^{0^*} &= S^{0^*} = \frac{1}{\sqrt{6}} . \end{aligned} \tag{3.11}$$

where P^{ω^*} indicates the symmetric fixed point value of an order parameter P corresponding to a rotation rate ω. The fixed point for a changing task can then be written as:

$$\begin{aligned} Q^{\omega^*} &= Q^{0^*} + \omega\, q^* , \\ C^{\omega^*} &= C^{0^*} + \omega\, c^* , \\ R^{\omega^*} &= R^{0^*} + \omega\, r^* , \\ S^{\omega^*} &= S^{0^*} + \omega\, s^* . \end{aligned} \tag{3.12}$$

where ω indicates the frequency of the rotation and q^*, r^*, c^*, s^* are deviations from the symmetric fixed point.

The system of equations (3.10) is then in terms of q^*, r^*, c^*, s^*, can be linearised using Taylor series expansion in ω which is assumed to be very small. The new symmetric fixed point found corresponds to:

$$\begin{aligned} q^* &= c^* = 0 , \\ r^* &= s^* = \frac{-5\sqrt{10}\pi}{112\eta} . \end{aligned} \tag{3.13}$$

To check the validity of the assumptions made to find the symmetric fixed point, we compare the numerical and analytical values of: $\eta(\underbrace{R_{symmetric}}_{teacher\ rotating} - \underbrace{R^{analytical}_{symmetric}}_{teacher\ fixed})$ for a given η and for different values of ω. The whole set of equations is solved numerically. We notice in figure 3.4 that the numerical and analytical solutions are equal for $\omega = 0$ and remain very close for larger rotation rates validating the analytical work. The difference observed results certainly from the fact that the original system is first modified by assuming that student vectors are in the teacher space, and then it is linearised.

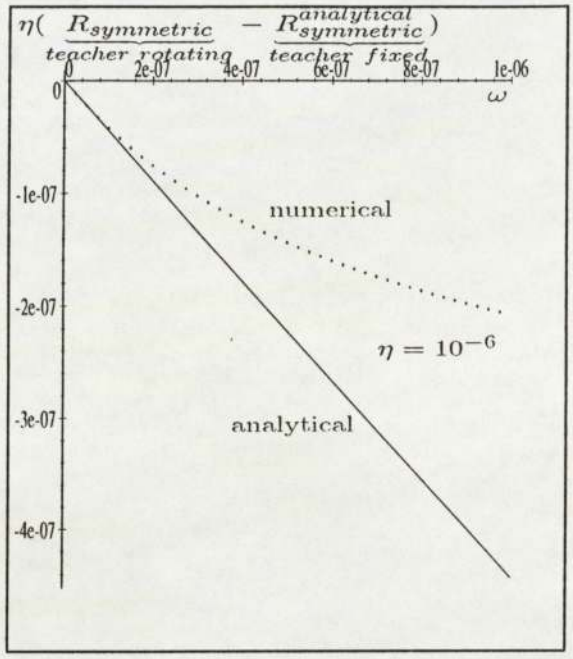


Figure 3.4: Numerical and analytical behaviour of $\eta($ $\underbrace{R_{symmetric}}_{teacher\ rotating}$ $-$ $\underbrace{R^{analytical}_{symmetric}}_{teacher\ fixed}$ $)$. The numerical and analytical values are very close to one another which validate the the expression of the analytical symmetric fixed point.

So far, the fixed point has been found under certain assumptions. The next stage is to study the escape from the symmetric plateau. To gain insight for the escape from the symmetric phase in the non-stationary environment we will first review the mechanism of this escape in the fixed environment framework. The student vectors, in that case, point towards the same direction (relative to teacher vectors) and the onset of the escape is characterised by a change in this direction in an attempt to imitate teacher vectors. The order parameters are as follows:

$$\begin{aligned} Q^{\omega} &= Q^{\omega^*} + q \,, \\ C^{\omega} &= C^{\omega^*} + c \,, \\ R^{\omega} &= R^{\omega^*} + r \,, \\ S^{\omega} &= S^{\omega^*} + s \,. \end{aligned} \tag{3.14}$$

where ω indicates the frequency of the rotation and q, r, c, s are deviations from the symmetric fixed point. In figure (3.1), it can be noticed that r and s escape before q and c. By assuming that student vectors lie in the teacher space it is possible to write: $Q = R^2 + S^2$ and $C = 2\ R\ S$ and an expansion around the asymptotic values gives the results: $Q = Q^* + 2R^*(r+s)$ and $C = C^* + 2R^*(r+s)$, so $q = c$ in the onset of the escape.

To reduce the number of variables further, we will make an assumption about the linear expansion of the generalisation error around the symmetric fixed point: $\epsilon_g(P) = \epsilon_g(P^*) + \underbrace{\nabla\epsilon_g(P - P^*)}_{\approx 0}$ (P denoting the system parameters). This is motivated by figure 3.2 where we can notice that a small deviation from the symmetric fixed point should not influence a lot the value of the generalisation error. The expression of the generalisation error is given in equation 2.11 , and it can be linearised by assuming a small $\frac{\omega}{\eta}$ ratio. The relation above implies a simple relation between the parameters:

$$q = \frac{168\sqrt{10}\ (r+s) - 5\sqrt{15}\ \pi\frac{\omega}{\eta}(5\sqrt{15} + r + s)}{12(14\sqrt{15} - 5\frac{\omega}{\eta}\pi)}\ . \tag{3.15}$$

The method used to reduce the number of parameters relies both on the fact that student vectors are in the teacher space and that the generalisation error does not vary a lot around the symmetric fixed point. It would have been possible to use only the former assumption and replace the expressions of Q and C by $R^2 + S^2$ and $2\ RS$ respectively.

The variables q, r, c and s have then only two degrees of freedom as they are linked by two equations. As r and s seem to dominate the escape from the symmetric plateau, we choose them to represent the evolution of the system. The dynamics reduces to:

$$\begin{aligned}
\frac{dr}{dt} &= -\Big(\frac{8\sqrt{15}\ \eta}{75\ \pi} + \frac{11}{70}\ \omega\Big)\ r - \Big(\frac{2\sqrt{15}\ \eta}{15\ \pi} + \frac{15}{28}\ \omega\Big)\ s\ , \\
\frac{ds}{dt} &= -\Big(\frac{2\sqrt{15}\ \eta}{15\ \pi} + \frac{15}{28}\ \omega\Big)\ r - \Big(\frac{8\sqrt{15}\ \eta}{75\ \pi} + \frac{11}{70}\ \omega\Big)\ s\ .
\end{aligned} \tag{3.16}$$

This is a set of linear differential equations whose matrix is real and symmetric. It has then two real eigenvalues and a general solution of the equation (for two different

eigenvalues) has the form: $\phi_1 \, e^{\lambda_1 t} + \phi_2 \, e^{\lambda_2 t}$ where ϕ_1 and ϕ_2 are the eigenvectors associated to the eigenvalues λ_1, λ_2 of the matrix. An escape from the symmetric phase can be observed if one of the eigenvalues is positive, as negative eigenvalues suppress the related exponential. Therefore, the only interesting eigenvalue, in this context, is the positive one which corresponds to :

$$\lambda_+ = \underbrace{\frac{4\sqrt{15}}{75\,\pi}\,\eta}_{\textit{fixed environment}} + \frac{53}{70}\,|\omega| \,. \tag{3.17}$$

The absolute value of the rotation angle here results from an invariance to the direction of rotation.

In order to check the validity of the assumptions made to reduce the number of variables, we study the positive eigenvalue of the linearised system of equation with its four variables: q, r, c and s and compare it to what is found for the reduced system. The new parts of the positive eigenvalue are compared to avoid scaling problems and the result is shown in figure 3.5. We notice that the curves are very close to one another for small values of ω which validate the method used to reduce the dimension.

The expression found for the positive eigenvalue is the most important result in this section. It shows that the positive eigenvalue which leads to escape from the symmetric phase is larger for the non-stationary case than for the fixed environment case. The escape is therefore much quicker as it is an exponential in λ (even if ω is small). Although this result relies on the small η approximation it is carried over for higher learning rates as observed in figure 3.1.

Summary of the symmetric phase

The symmetric phase is characterised by a fixed point close to the one corresponding to a stationary task, and the length of the symmetric plateau is shortened significantly compared to the fixed environment. This important result is shown for a modification of the task represented by rotations, and by assuming that $\frac{\omega}{\eta}$ and η are small. This last assumption has two implications used in the previous calculations, the first is that

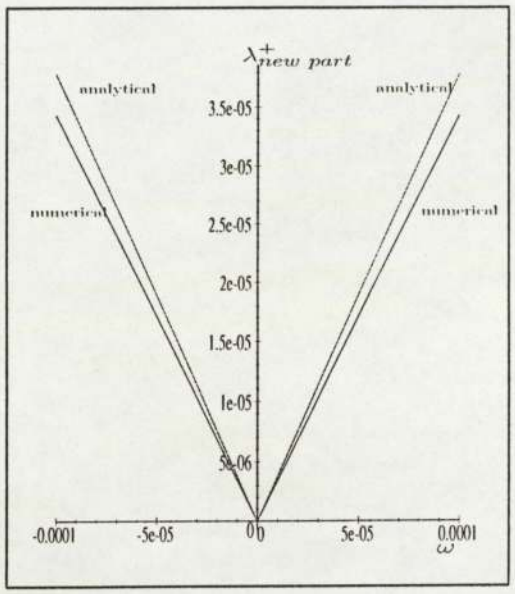


Figure 3.5: Comparison of the analytical and numerical solutions for the new part of the positive eigenvalue. The curves are close to one another and therefore the assumptions made to reduce the dynamics are valid.

ω is also small and the second concerns student vectors ($\mathbf{J}_1$, $\mathbf{J}_2$) who are confined to the teacher space.

The study of the eigenvalue of the escape gave insight to suggesting a better training algorithm as described in chapter 4.

3.2.2 Convergence phase

In this second phase of the learning process we examine the final convergence of the model to the asymptotic fixed point. This was carried out in [3] for the fixed environment case where each student node specialises on a specific teacher node. In figure 3.2 we show the evolution of the generalisation error; it is very difficult to gain insight about the effect of non-stationarity on the convergence phase because its scaling is much smaller than the one for the symmetric phase. Zooming on the tail of convergence (figure 3.6) reveals a residual error for a non stationary task i.e. learning is then not perfect asymptotically.

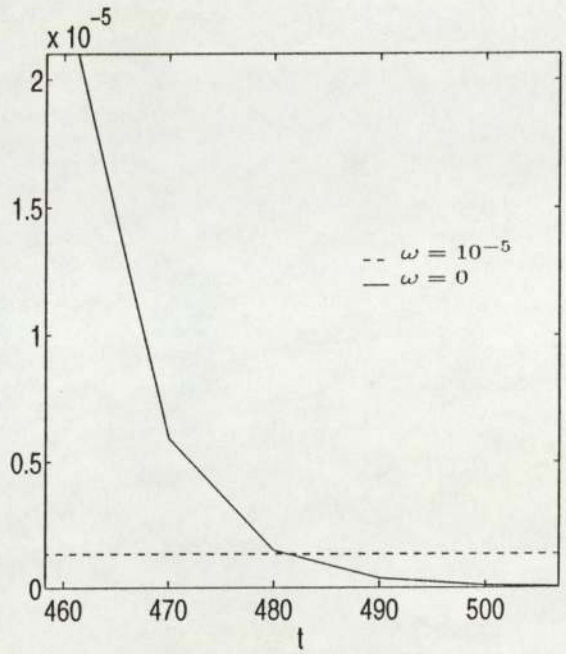


Figure 3.6: Generalisation error for $\omega = 10^{-5}$ and $\omega = 0$ ($\eta = 1.66$). The curve is a zoom on the convergence phase showing a residual error for the case $\omega \neq 0$.

The convergence to the asymptotic values depends critically on the learning rate η which we no longer require to be small. The whole system can be represented by seven equations of seven variables (Appendix C). The first important feature of the convergence phase is the asymptotic fixed point obtained by solving the system of equations:

$$\frac{dR_{in}}{dt} = 0 \quad \forall i, n \qquad \frac{dQ_{ik}}{dt} = 0 \quad \forall i, k \; . \tag{3.18}$$

By solving the system of equations numerically for different values of ω and for a fixed η, we obtain the curves shown in figure 3.7.

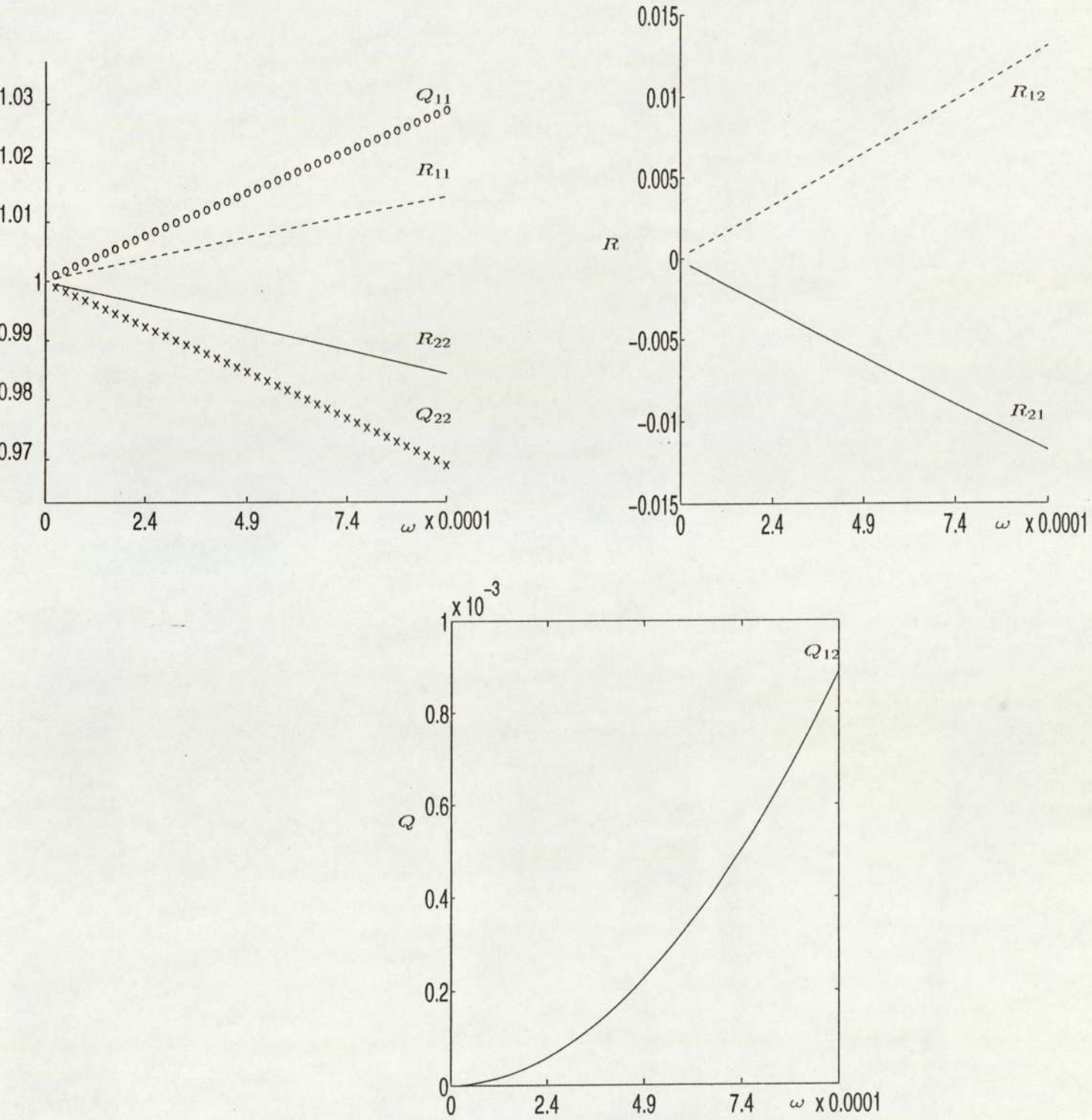


Figure 3.7: Numerical dependence of the asymptotic values of the order parameters on ω (η=0.5).

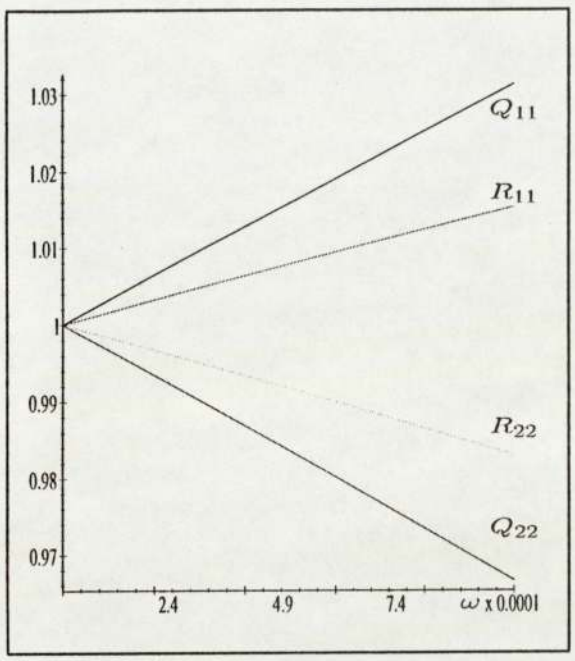
The most important result shown by these curves is that Q_{12} is quadratic in terms of ω unlike the other order parameters which are linear. By supposing that the frequency of the teacher rotation is small, the asymptotic point will be close to the one for a fixed task. Therefore, we expand the order parameters around the stationary case fixed point as follows:

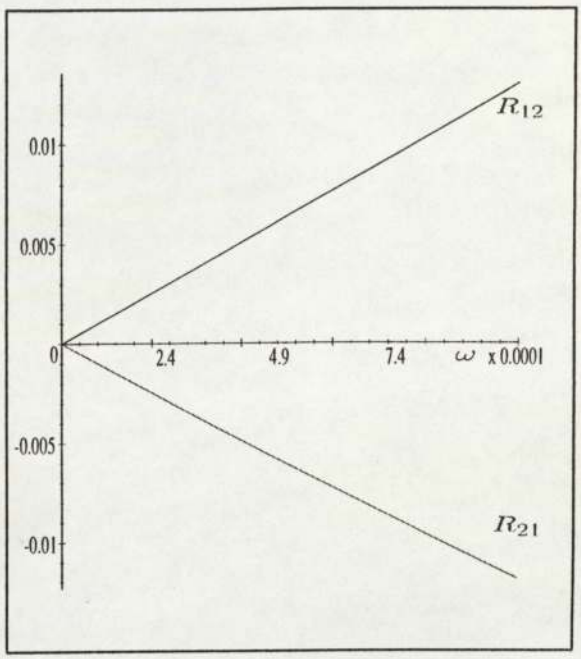
$$\begin{aligned} R_{11} = 1 + r_{11},\ R_{12} = r_{12},\ R_{21} = r_{21},\ R_{22} = 1 + r_{22} \\ Q_{11} = 1 + q_{11},\ Q_{12} = q_{12},\ Q_{22} = 1 + q_{22}\ . \end{aligned} \tag{3.19}$$

By linearising the system of equations to first order in q_{11}, q_{12}, q_{22}, r_{11},r_{12}, r_{21}, r_{22} and then by expanding the solution of the linear system to second order in ω, one can obtain approximated expressions for the asymptotic fixed point. The expressions found are complicated and given in Appendix D. By plotting the curves showing the dependency on ω for the analytical asymptotic point, it is easy to notice that they are similar to those found numerically (figure 3.8).

The asymptotic fixed point depends on both ω and η, so when the learning rate is modified the asymptotic point is modified as well. In order to study this, the generalisation error is calculated for this asymptotic point and its behaviour in terms of the two variables η and ω is presented in figure 3.9.

This curve shows first that the generalisation error is diverging for particular values $\eta \geq 2.3$. So when $\eta > 2.3$ defining the maximal learning rate which seems to be dependent of ω. Another method based on an eigenvalue analysis giving the same maximal learning rate is examined later on. When $\eta < 2.3$, the error may have a non zero minimum with respect to the learning rate used for certain values of ω. Some contours obtained by solving numerically the original system of equations 3.9 are drawn for different values of ω in figure 3.10.





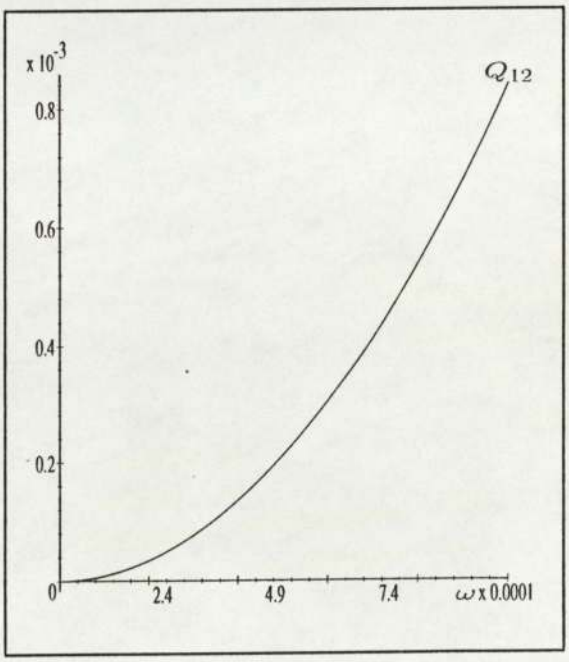


Figure 3.8: Analytical dependence of the asymptotic values of the order parameters on ω (η=0.5).

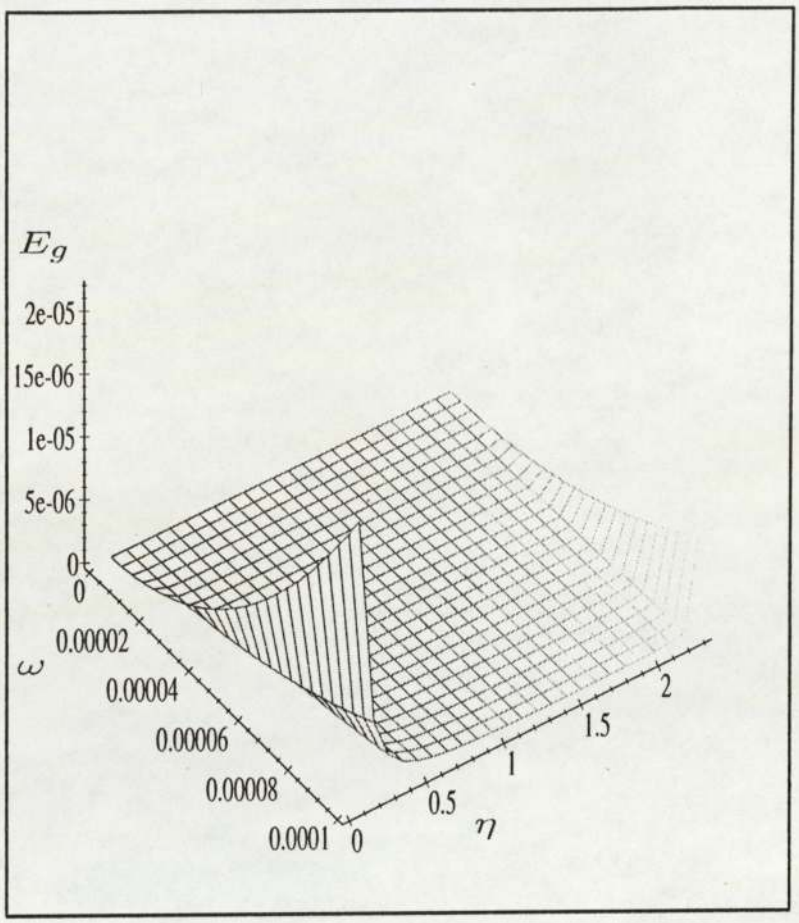


Figure 3.9: Generalisation error at the asymptotic fixed point in terms of ω and η

The optimal learning rate (denoted η_{opt_err}) minimising the error seems to depend on ω. Moreover by observing the curves in figure 3.10, it appears that it is increasing with the rotation rate. A numerical study of this property enables to plot figure 3.11 implying that the optimal learning rate grows with the rotation rate.

The learning rate η_{opt_err} minimises the generalisation error disregarding the speed of convergence. Before investigating the learning rate optimising the speed of convergence, we will examine the residual error.

As I mentioned before, the asymptotic generalisation error remains non zero. To explain the residual error, we look at the projection of the first student vector $\mathbf{J}_1$ on

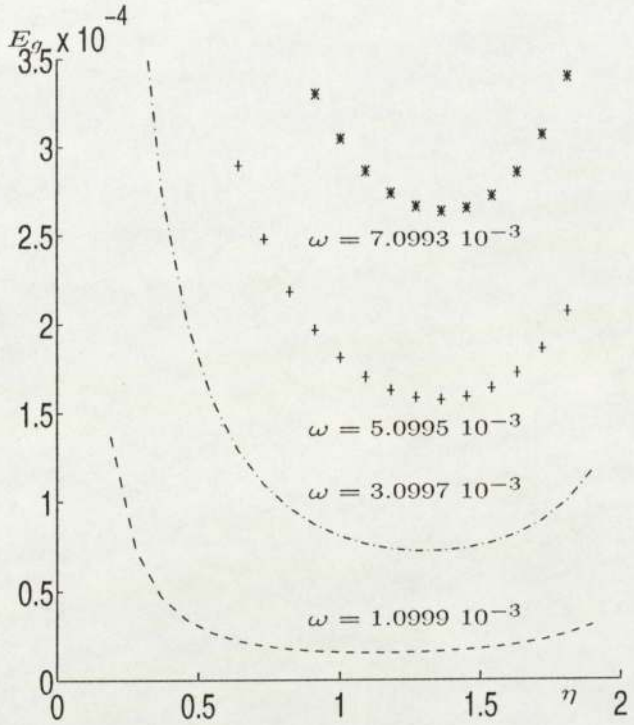


Figure 3.10: The generalisation error increases according to ω. The error seems to be minimised by a certain learning rate for each given ω.

the various teacher vectors. The angle θ between $\mathbf{J}_1$ and $\mathbf{B}_1$ can be expressed as:

$$\theta = \arctan\left(\frac{R_{12}}{R_{11}}\right), \tag{3.20}$$

and remains non zero even when the process converges (which explains the residual error). This is only one projection and there are many others; we then assume a residual constant "phase shift" between the dynamics of the teacher and that of the student which depends on the relation between the learning rate η and the rotation rate ω. The angle θ is then studied numerically for different values of ω and the result is shown in figure 3.12.

The dependence of θ on ω is linear implying that a bigger "phase shift" between the teacher and the student results from a bigger rotation rate.

So far, the discussion focused on the minimum of the asymptotic error and not on the speed of convergence to a given asymptotic fixed point. Now we study the convergence speed by looking at the dynamics of a set of vectors representing deviations

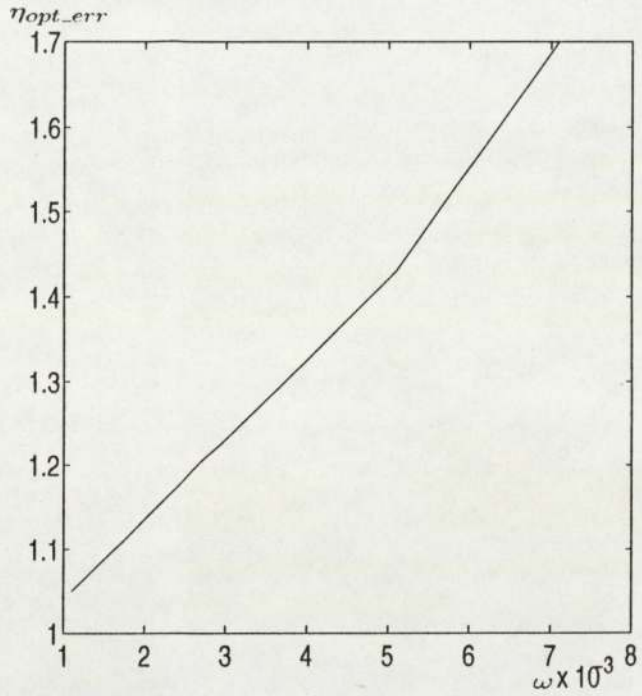


Figure 3.11: The learning rate optimising the generalisation error increases according to ω.

from the asymptotic fixed point:

$$\frac{d}{dt}\begin{pmatrix} Q_{11}-Q_{11}^* \\ Q_{12}-Q_{12}^* \\ Q_{22}-Q_{22}^* \\ R_{11}-R_{11}^* \\ R_{12}-R_{12}^* \\ R_{21}-R_{21}^* \\ R_{22}-R_{22}^* \end{pmatrix} = \mathcal{M}\begin{pmatrix} Q_{11}-Q_{11}^* \\ Q_{12}-Q_{12}^* \\ Q_{22}-Q_{22}^* \\ R_{11}-R_{11}^* \\ R_{12}-R_{12}^* \\ R_{21}-R_{21}^* \\ R_{22}-R_{22}^* \end{pmatrix}, \tag{3.21}$$

where $\mathcal{M}$ is the 7×7 matrix describing the system and P^* indicates the asymptotic value of order parameter P. There are normally seven eigenvalues for $\mathcal{M}$ (equation 3.21), and they are negative in a relevant domain for η when ω is maintained constant. Among the 7 eigenvalues there are only two which dominate the dynamics on the long-time corresponding to the largest values and their dependence on η is presented in figure 3.13.

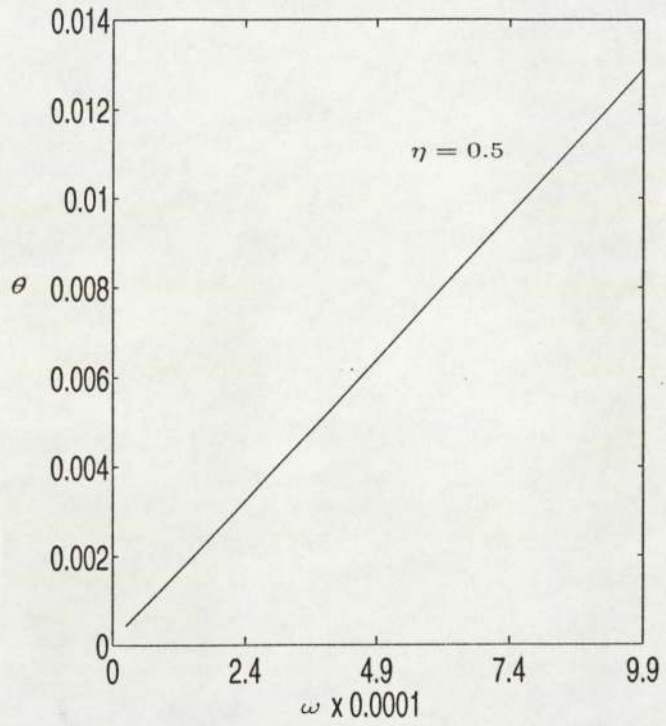


Figure 3.12: Dependence of the angle made by the projection of $\mathbf{J}_1$ in the teacher space with vector $\mathbf{B}_1$ for $K = M = 2$. The curve is a straight line implying a linear dependence.

The eigenvalue λ_1 is a non-linear function of η (for a fixed ω) and negative for small η. The eigenvalue λ_2 is linear in η. For large η, λ_1 becomes positive and training does not converge to the optimal solution, defining the maximum learning rate η_{max} as $\lambda_1(\eta_{max}) = 0$. The value found for η_{max} corresponds to that observed in figure 3.9.

In order to identify the convergence time τ, which is inversely proportional to the modulus of the eigenvalues associated with the slowest decay mode, we expand the generalisation error to second order in our parameters. We find that the mode associated with the linear eigenvalue does not contribute to first order terms, and when η is small controls only second order terms with a decay rate of $2\lambda_2$. The learning rate η_{opt_conv}, providing the fastest asymptotic decay rate of the generalisation error, is therefore either given by the condition $\lambda_1(\eta_{opt\_conv}) = 2\ \lambda_2(\eta_{opt\_conv})$ or alternatively by $min_\eta(\lambda_1)$ if $\lambda_1 > 2\lambda_2$ at the minimum of λ_1.

By observing figure 3.13, we can then say that the optimal learning rate corresponds

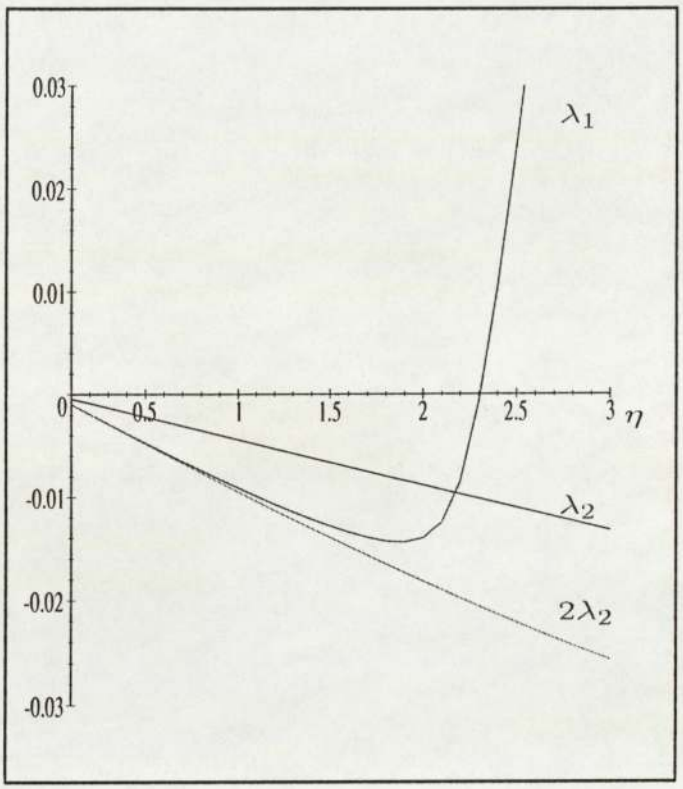


Figure 3.13: Dependence of the two most important eigenvalues describing the convergence to the asymptotic values when $0 < \eta < 3$ for a fixed ω=10^{-4}. The system corresponds to $K = M = 2$ case.

to the minimum of λ_1 (around $\eta \approx 1.8$) and is bigger than η_{opt_err} (figure 3.11). This comparison shows that this two learning rates are not similar and have different values for at least one specific rotation rate.

The optimal choice for the learning rate could be to start at the fastest convergence η and then switch to the lowest error learning rate.

3.2.3 Summary of the learning process

Studying $K = M = 2$ case reveals that the symmetric phase is significantly shorter when learning non-stationary tasks as the eigenvalues of the system which determine the escape time are bigger than for a stationary environment. By assuming small rotation rate, the symmetric fixed point can be found by expanding around the fixed

point obtained previously for stationary tasks. This provides insight to suggest a better training algorithm shown in next chapter, which speeds up the escape from the symmetric phase.

The study of the convergence phase has been carried out mainly numerically because of the complexity of the system of equations. We have found the asymptotic values of the order parameters for fixed η, revealing a convergence to a sub-optimal fixed point. We examined numerically the generalisation error as a function of the learning rate η for fixed values of ω obtaining different minima with respect to η. This result implies an ω dependence of η_{opt_err} which is the learning rate minimising the generalisation error for a given ω.

The speed of convergence is also investigated. We find another optimal learning rate η_{opt_conv} which minimises the time needed to reach a given asymptotic fixed point, by studying the eigenvalues of the dynamics.

For each rotation rate ω, there are at least two optimal choices for η related to the convergence phase. One leading to an optimal value of the generalisation error and the other leading to the asymptotic fixed point very quickly.

As convergence is imperfect, the residual error is studied by projecting one student vector on the corresponding teacher vector, finding a "phase shift" between the student and the teacher orthonormal basis i.e., the angle made by the student and the teacher vector learned by it remains a non zero constant depending linearly on the rotation rate ω.

Because of time limitations some investigations are left unfinished, such as the dependence of η_{opt_conv} on ω and the study of the general case of simultaneous rotations in several directions when any number of hidden nodes may be used.

Chapter 4

Modified back-propagation

When analysing the symmetric phase in the previous chapter we have noticed that the length of the symmetric plateau is shortened significantly by considering a permanent rotation of the teacher vectors. We would like to use the insight gained for modifying the dynamics in the fixed environment case in such a way that will speed up the escape from the symmetric phase. This idea is motivated by the behaviour in the $K = M = 2$ case:

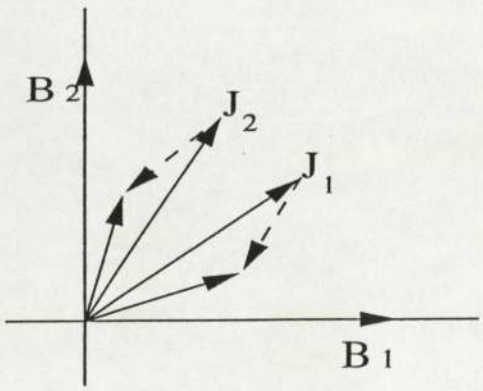


Figure 4.1: Teacher vectors ($\mathbf{B}_1$, $\mathbf{B}_2$) are orthonormal and remain fixed. Subtracting a part of the student $\mathbf{J}_2$ to $\mathbf{J}_1$ makes it farther from $\mathbf{J}_2$.

Subtracting from each student vector a part of the other vector may assist in breaking the symmetric phase by separating student vectors which are pointing almost in the same direction. This way students are attracted to different directions, increasing

their correlation with different teacher vectors. Once the system has started converging this additional term is not useful anymore and should be switched off to allow for the convergence of the student network. The aim of this chapter is to analyse the system with the modified algorithm and to define to what extent it improves the training performance.

The new learning rule can then be expressed as follows:

$$\mathbf{J}_i^{t+1} = \mathbf{J}_i^t + \frac{\eta}{N}\,\delta_i^t \boldsymbol{\xi}^t - \frac{\gamma}{N}\sum_{k\neq i}\mathbf{J}_k^t\,. \tag{4.1}$$

The same order in N is kept for both the learning rate and the coefficient γ to keep the dynamics smooth. The study of a non-stationary environment has been carried out by considering a simple case as it was the building block for the general one; in this case there is no need to restrict the study and the focus is immediately on the general configuration. However, one simplification will be taken by supposing that $K = M$. This case is amenable to analysis and corresponds to a realisable scenario [3] where the number of student nodes is the same as the number of teacher nodes.

4.1 Dynamics for the learning process

The environment is supposed to be fixed and teacher vectors are denoted $\mathbf{B}_n$. The teacher-student correlation, at a time t is given by: $R_{in}^t = \mathbf{J}_i^t \,.\, \mathbf{B}_n$, therefore:

$$\begin{aligned} R_{in}^{t+1} &= \mathbf{J}_i^{t+1} \,.\, \mathbf{B}_n \\ &= (\mathbf{J}_i^t + \frac{\eta}{N}\,\delta_i^t\boldsymbol{\xi}^t - \frac{\gamma}{N}\sum_{k\neq i}\mathbf{J}_k^t) \,.\, \mathbf{B}_n \\ &= R_{in}^t + \frac{\eta}{N}\,\delta_i^t\boldsymbol{\xi}^t \,.\, \mathbf{B}_n - \frac{\gamma}{N}\sum_{k\neq i}R_{kn}^t\,, \end{aligned} \tag{4.2}$$

so by considering two consecutive time steps:

$$\frac{R_{in}^{t+1} - R_{in}^t}{1/N} = \eta\;\delta_i^t y_n^t - \underbrace{\gamma\sum_{k\neq i}R_{kn}^t}_{O(1)}\,. \tag{4.3}$$

As done in the previous chapters, averages through the activations of the nodes and thermodynamic limit are performed to get continuous differential equations. For

teacher-student overlaps the dynamics are as follows:

$$\frac{dR_{in}}{dt} = \eta \, \langle \delta_i y_n \rangle - \gamma \sum_{k \neq i} R_{kn}^t \, . \tag{4.4}$$

The equation for the student-student overlaps can be written similarly as:

$$\begin{aligned} Q_{ij}^{t+1} &= (\mathbf{J}_i^t + \frac{\eta}{N}\delta_i^t \boldsymbol{\xi}^t - \frac{\gamma}{N}\sum_{q\neq i}\mathbf{J}_q^t) \cdot (\mathbf{J}_j^t + \frac{\eta}{N}\delta_j^t \boldsymbol{\xi}^t - \frac{\gamma}{N}\sum_{p\neq j}^{N}\mathbf{J}_p^t) \\ &= Q_{ij}^t + \frac{\eta}{N}(\delta_i^t x_j^t + \delta_j^t x_i^t) - \frac{\gamma}{N}(\sum_{p\neq j} Q_{pi}^t + \sum_{q\neq i} Q_{qj}^t) \\ &- \frac{\eta\gamma}{N^2}(\sum_{q\neq i} x_q^t \delta_j^t + \sum_{p\neq j} x_p^t \delta_i^t) + \frac{\gamma^2}{N^2}\sum_{q\neq i}\mathbf{J}_q^t \cdot \sum_{p\neq j}\mathbf{J}_p^t + \frac{\eta^2}{N^2}\delta_i^t\delta_j^t \, \boldsymbol{\xi}^t \cdot \boldsymbol{\xi}^t \, . \end{aligned} \tag{4.5}$$

After considering the averages of the fluctuating quantities and by examining the order of the expressions:

$$\begin{aligned} \frac{Q_{ij}^{t+1} - Q_{ij}^t}{1/N} &= \underbrace{\eta \, \langle \delta_i^t x_j^t + \delta_j^t x_i^t \rangle - \gamma \, \langle \sum_{p\neq j} Q_{pi}^t + \sum_{q\neq i} Q_{qj}^t \rangle}_{O(1)} \\ &- \underbrace{\frac{\eta\gamma}{N}\, \langle \sum_{q\neq i} x_q^t \delta_i^t + \sum_{p\neq i} x_p^t \delta_k^t \rangle + \frac{\gamma^2}{N}\, \langle \sum_{q\neq i}\mathbf{J}_q^t \sum_{p\neq j}\mathbf{J}_p^t \rangle}_{O(\frac{1}{N})} \\ &+ \underbrace{\frac{\eta^2}{N}\, \langle \delta_i^t \delta_j^t \, \boldsymbol{\xi}^t \cdot \boldsymbol{\xi}^t \rangle}_{O(1)} \, . \end{aligned} \tag{4.6}$$

The last element is $O(1)$ because it corresponds to the variance of the distribution of the inputs which is a normal Gaussian (zero mean and unit variance) for each component of $\boldsymbol{\xi}$. The thermodynamic limit simplifies the expression above as terms of $O(\frac{1}{N})$ are neglected and the continuous differential equations is given by:

$$\frac{dQ_{ij}}{dt} = \eta \, \langle \delta_i x_j + \delta_j x_i \rangle - \gamma \, (\sum_{p\neq h} Q_{pi} + \sum_{q\neq i} Q_{qj}) + \eta^2 \, \langle \delta_i \delta_j \rangle \, .$$

The new dynamics can then be expressed in terms of those for a gradient descent as follow (P^γ indicates the order parameter P for the modified back propagation having a coefficient γ):

$$\begin{aligned} \frac{dQ_{hi}^\gamma}{dt} &= \frac{dQ_{hi}^0}{dt} - \gamma \, (\sum_{p\neq h} Q_{pi} + \sum_{q\neq i} Q_{qh}) \, , \\ \frac{dR_{in}^\gamma}{dt} &= \frac{dR_{in}^0}{dt} - \gamma \sum_{k\neq i} R_{kn} \, . \end{aligned} \tag{4.7}$$

The elements $\frac{dP^0}{dt}$ are defined in equation 2.17.

Numerical solutions of the system are carried out to observe the evolution of the learning process. To have a general idea, we look at the case of $K = M = 3$. To compare the learning process with and without the new term, it is useful to find the solutions of the $K = M = 3$ dynamics in the original framework (equation 2.5). The numerical behaviour of the dynamics without a new term for the back-propagation is shown in figure 4.2.

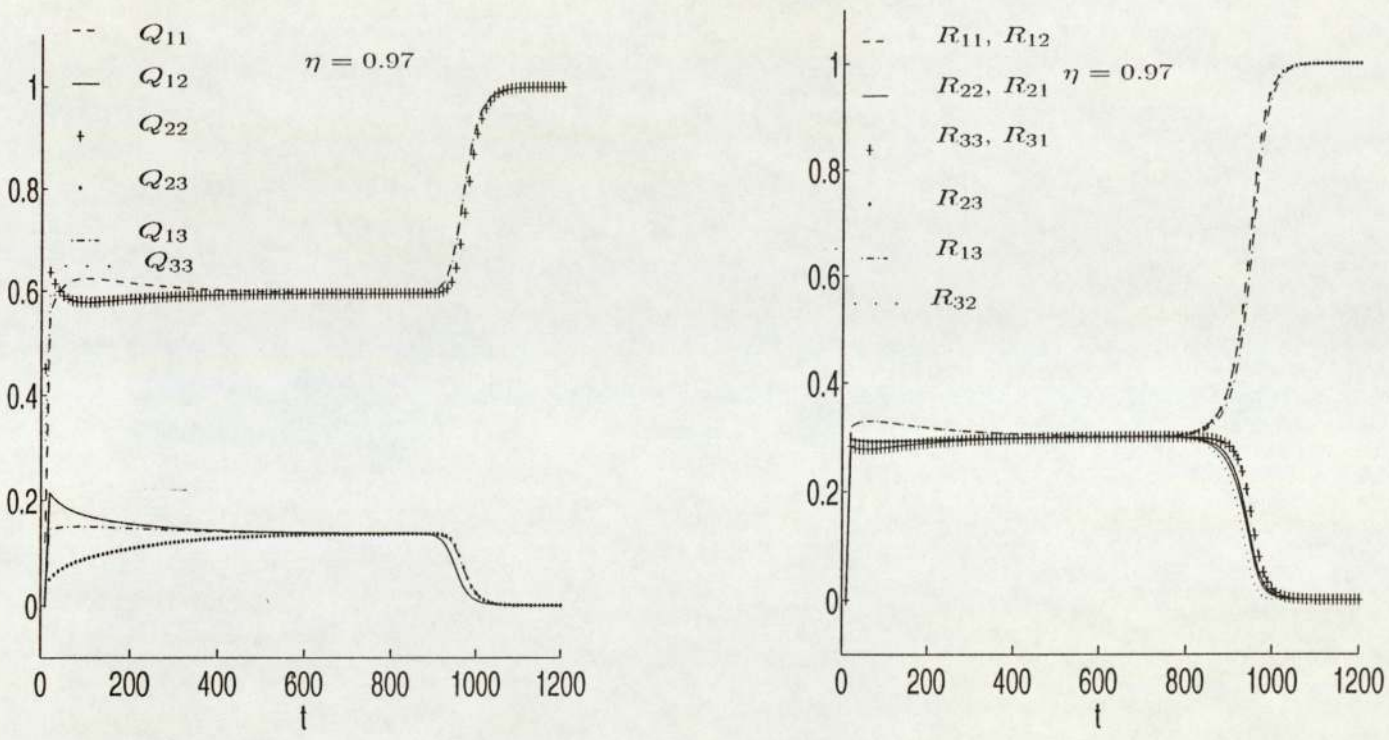


Figure 4.2: Dynamical evolutions of the order parameters for $\gamma = 0$ corresponding to gradient descent ($\eta = 0.97$) and for a $K = M = 3$ case.

The behaviour for the modified back-propagation is similar to the one for the gradient descent (figure 2.2) as observed in figure 4.3.

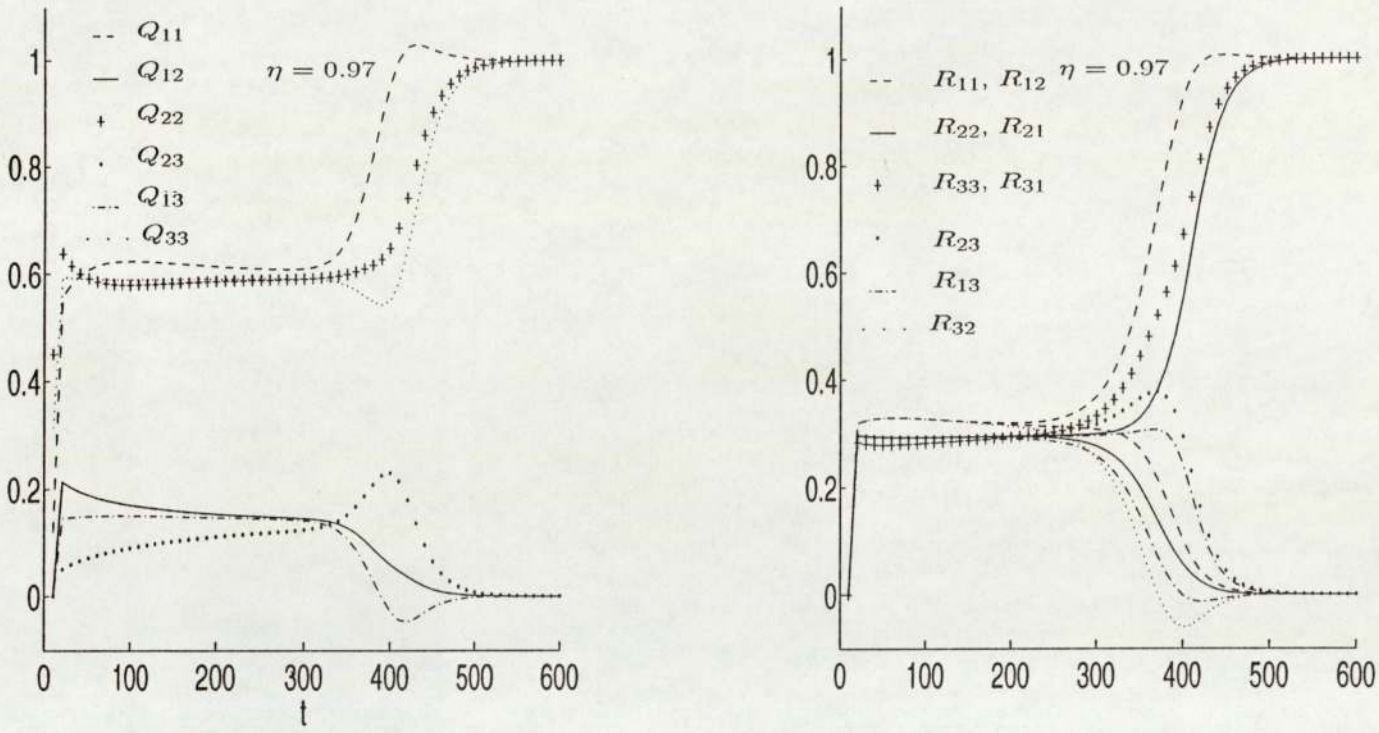


Figure 4.3: Dynamical evolutions of the order parameters for $\gamma = 10^{-5}$ corresponding to the modified gradient descent ($\eta = 0.97$) and for a $K = M = 3$ case. The graphs are qualitatively similar to those observed for $\gamma = 0$.

As mentioned before we focus here only on the symmetric phase. The length of the symmetric plateau has been shortened from 1000 when $\gamma = 0$ to 400 when $\gamma = 10^{-5}$. The reduction is emphasised in the graph for the generalisation error for both $\gamma = 0$ and $\gamma = 10^{-5}$ (figure 4.4).

This behaviour for small values of γ seems to agree with the observation made in the beginning of this chapter and with the intuition which has motivated the study.

An analytical study is then performed to explain the reduction of the length of the symmetric plateau and its dependence on the coefficient γ.

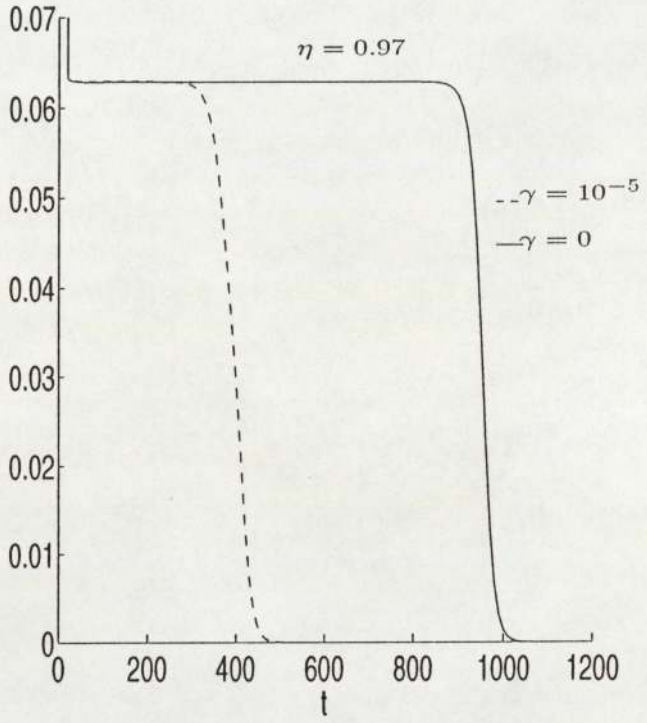


Figure 4.4: Generalisation error for $\gamma = 10^{-5}$ corresponding to the modified gradient descent ($\eta = 0.97$) and for a $K = M = 3$ case. The length of the symmetric plateau becomes shorter for $\gamma = 10^{-5}$.

4.2 Analysis of the symmetric phase

As in section 3.2.1, the symmetric fixed point has to be calculated first. Initially there are $\frac{K(K+1)}{2}$ student-student overlaps and K^2 teacher-student overlaps, however in figure 4.3 it is possible to see that all R_{in} have nearly the same numerical value. This should be coupled with the fact that we should keep the differentiation between R_{ii} and R_{in} ($i \neq n$) as it is important for the escape from the plateau. So, the R's will be described by two variables R and S as follows: $R_{in} = R\ \delta_{in} + (1 - \delta_{in})\ S$. The initial teacher-student overlaps are reduced then to two only on the basis of the numerical behaviour of the system.

Student-student overlaps have a similar behaviour except for the gap observed, in figure 4.3, between Q_{ij}, $i \neq j$ and Q_{ii}, and which vanishes for small η when $\gamma = 0$ [3]. They can be written as: $Q_{ij} = Q\ \delta_{ij} + (1 - \delta_{ij})\ C$, and then be represented by 2 variables only. The symmetric fixed point is then described by 4 variables Q^*, R^*, C^*, S^* which

makes the system of equations below equal to zero:

$$
\begin{aligned}
\frac{dQ}{dt} &= \frac{4\eta}{\pi(1+Q)}\left(\frac{R}{\sqrt{2(1+Q)-R^2}}+\frac{(K-1)S}{\sqrt{2(1+Q)-S^2}}-\frac{Q}{\sqrt{1+2Q}}-\frac{(K-1)C}{\sqrt{(1+Q)^2-C^2}}\right)\\
&- 2\,(K-1)\,\gamma\,C\,,\\
\frac{dR}{dt} &= \frac{2\eta}{\pi(1+Q)}\left(\frac{1+Q-R^2}{\sqrt{2(1+Q)-R^2}}-\frac{(K-1)RS}{\sqrt{2(1+Q)-S^2}}-\frac{R}{\sqrt{1+2Q}}\right.\\
&- \left.\frac{(K-1)S(1+Q)-RC(K-1)}{\sqrt{(1+Q)^2-C^2}}\right)-\gamma\,(K-1)\,S\,,\\
\frac{dS}{dt} &= \frac{2\eta}{\pi(1+Q)}\left(\frac{1+Q-(K-1)S^2}{\sqrt{2(1+Q)-S^2}}-\frac{SR}{\sqrt{2(1+Q)-R^2}}-\frac{S}{\sqrt{1+2Q}}\right.\\
&- \left.\frac{(R+S(K-2))(1+Q)-CS(K-1)}{\sqrt{(1+Q)^2-C^2}}\right)-\gamma\,(R+S(K-2))\,,\\
\frac{dC}{dt} &= \frac{4\eta}{\pi(1+Q)}\left(\frac{S(1+Q)-CR}{\sqrt{2(1+Q)-R^2}}+\frac{R(1+Q)+S(1+Q)(K-2)-CS(K-1)}{\sqrt{2(1+Q)-S^2}}\right.\\
&- \left.\frac{(1+Q)(Q+C(K-2)-C^2(K-1)}{\sqrt{(1+Q)^2-C^2}}-\frac{C}{\sqrt{1+2Q}}\right)-2\,\gamma\,(Q+C(K-2))\,.
\end{aligned}
\tag{4.8}
$$

As in section 3.2.1, this system is not linear in terms of Q, R, S, C. To linearise it, the coefficient γ is assumed small, which is consistent with values found to work well in practice used (see figure 4.3). The new fixed point should then be close to the one found for a classical gradient descent [3]:

$$
\begin{aligned}
Q^{0^\star} &= C^{0^\star} = \frac{1}{2K-1}\,,\\
R^{0^\star} &= S^{0^\star} = \frac{1}{\sqrt{K(2K-1)}}\,.
\end{aligned}
\tag{4.9}
$$

The new fixed point can be described as an expansion around $Q^{0^\star}$, $C^{0^\star}$, $R^{0^\star}$, $S^{0^\star}$ with respect to γ. The variables used are as follows:

$$
\begin{aligned}
Q^{\gamma^\star} &= Q^{0^\star}+\gamma\,q^*\,,\\
C^{\gamma^\star} &= C^{0^\star}+\gamma\,c^*\,,\\
R^{\gamma^\star} &= R^{0^\star}+\gamma\,r^*\,,\\
S^{\gamma^\star} &= S^{0^\star}+\gamma\,s^*\,.
\end{aligned}
\tag{4.10}
$$

where P^γ indicates the value of the order parameter P for modified back-propagation

with a coefficient γ. The symmetric point is then given by:

$$\begin{aligned} q^* &= c^* = -\frac{(2K^2 - K - 1)\sqrt{2K+1}\pi}{K(2K-1)^{\frac{5}{2}}\eta}, \\ r^* &= s^* = -\frac{(2K^2 - K - 1)\sqrt{2K+1}\pi}{2K^{\frac{3}{2}}(2K-1)^2\eta}. \end{aligned} \tag{4.11}$$

In order to know the reliability of this analytical result, we compare the numerical and analytical values of: $\eta(\underbrace{R_{symmetric}}_{modified\ BP} - \underbrace{R^{analytical}_{symmetric}}_{classical\ BP})$ (BP indicates back-propagation) in the case $K = M = 3$ for a given η and for different values of γ. The set of equations is then solved numerically. We notice in figure 4.5 that the numerical and analytical solutions are equal for $\gamma = 0$ and remain very close which validate the analytical expression of the symmetric fixed point. The difference observed results certainly from the fact that the original system is first linearised then solved.

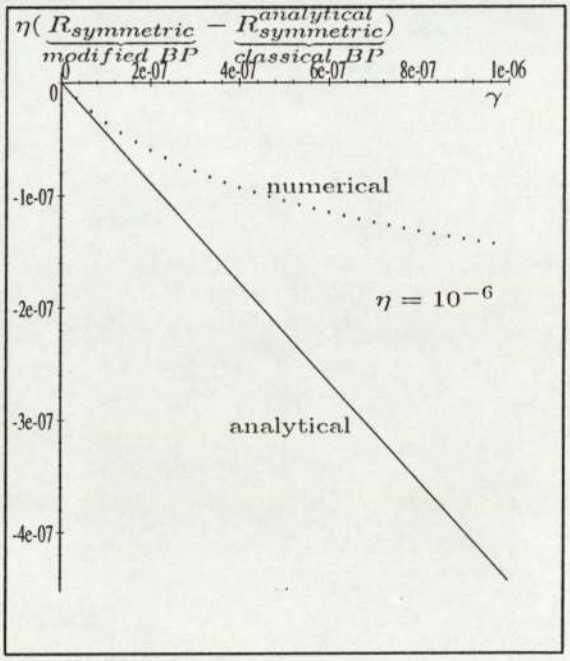


Figure 4.5: Numerical and analytical behaviour of $\eta(\underbrace{R_{symmetric}}_{modifed\ BP} - \underbrace{R^{analytical}_{symmetric}}_{classical\ BP})$ for the case $K = M = 3$ (BP indicates back-propagation). The numerical and analytical values are very close to one another which validate the expression found for the analytical symmetric fixed point.

We will now look at the escape from the symmetric phase. The order parameters

are then as follows:

$$\begin{aligned} Q^\gamma &= Q^{\gamma^*} + q\,, \\ C^\gamma &= C^{\gamma^*} + c\,, \\ R^\gamma &= R^{\gamma^*} + r\,, \\ S^\gamma &= S^{\gamma^*} + s\,. \end{aligned} \tag{4.12}$$

By expanding to first order in q, r, c, s the expressions of Q and C ($Q = R^2 + S^2$ and $C = 2RS$, as the teacher orthogonal space is neglected), it is possible to find that $q = c$ at the onset of the escape.

To reduce the number of variables further, we will make an assumption about the linear expansion of the generalisation error around the symmetric fixed point: $\epsilon_g(P)=\epsilon_g(P^*)+\underbrace{\nabla\epsilon_g(P-P^*)}_{\approx 0}$ (P denoting the system parameters). This is motivated by figure 4.4 where we can notice that a small deviation from the symmetric fixed point should not influence a lot the value of the generalisation error. The expression of the generalisation error is given in equation 2.11 , and it can be linearised by assuming a small $\frac{\gamma}{\eta}$ ratio. The relation above implies a simple relation between the parameters:

$$q = \frac{168\sqrt{10}\,(r+s) - 5\sqrt{15}\,\pi\frac{\omega}{\eta}(5\sqrt{15}+r+s)}{12(14\sqrt{15}-5\frac{\omega}{\eta}\pi)}\,. \tag{4.13}$$

The method used to reduce the number of parameters relies both on the fact that student vectors are in the teacher space and that the generalisation error does not vary a lot around the symmetric fixed point. It would have been possible to use only the former assumption and replace the expressions of Q and C by $R^2 + S^2$ and 2 RS respectively.

The initial four parameters q, r, c and s have therefore only two degrees of freedom as they are linked by two equations. Then the dynamics can be written as a linear system in r and s by keeping first order terms in γ in the Taylor expansions:

$$\frac{d}{dt}\begin{pmatrix} r \\ s \end{pmatrix} = \left[-\eta K(2K-1)^{\frac{3}{2}}\,\mathcal{G} - \gamma\,\sqrt{2K+1}\,\mathcal{H}\right]\begin{pmatrix} r \\ s \end{pmatrix} \tag{4.14}$$

where:

$$\mathcal{G} = \begin{pmatrix} 2(4K^2 - 5K + 2) & 4(K-1)(2K^2 - 2K + 1) \\ 4(K-1)(2K^2 - 2K + 1) & 2(4K^3 - 8K^2 + 5K - 2) \end{pmatrix} \tag{4.15}$$

and

$$\mathcal{H} = \begin{pmatrix} (K-1)(\pi(K-1)(4K^2+ & (K-1)(\pi(8K^4 + 16K^3 - 60K^2+ \\ \quad 26K - 15) + 12K - 6) & \quad 54K - 15) + 6(K-1)(2K-1)) \\ & \\ (\pi(8K^4 + 16K^3 - 60K^2+ & (\pi(8K^5 + 4K^4 - 74K^3 + 111K^2+ \\ +54K - 15) + 6(K-1)(2K-1)) & -67K + 15) + 12K^3 - 30K^2 + 24K - 6) \end{pmatrix}$$

The escape is described by the positive eigenvalue as it has been explained in section 3.2.1. Here, it is given in terms of the positive eigenvalues calculated for gradient descent and an additional term:

$$\lambda_+ = \lambda_+^0 + \lambda_+^\gamma = 2 \underbrace{\frac{\eta K}{\sqrt{2K-1}\pi(2K+1)^{\frac{3}{2}}}}_{back-propagation} + \frac{(4K^3 - 2K^2 + 3K - 2)\gamma}{(2K+1)(2K-1)^2} \tag{4.16}$$

The new part in the eigenvalue is positive as γ is taken positive by definition. This explains the shorter length of the symmetric plateau for the adaptive gradient descent shown in the numerical solutions (figure 4.4). This important result has been found under assumptions motivated by the numerical behaviour, i.e. $\eta, \frac{\gamma}{\eta}$ small implying that also γ is small. Although the result found relies on the small η approximation, it is carried over for larger learning rates as observed in figure 4.3.

To examine the validity of the assumptions we study the positive eigenvalue of the linearised system of equations with its four variables: q, r, c, s and compare it to what is found when the system is reduced. To avoid scaling problems, only λ_+^γ values are compared. The parameters chosen are consistent with the assumption presented above. The curve obtained is shown in figure 4.6.

The curves are close to one another which implies that the assumptions made for the analytical study are valid.

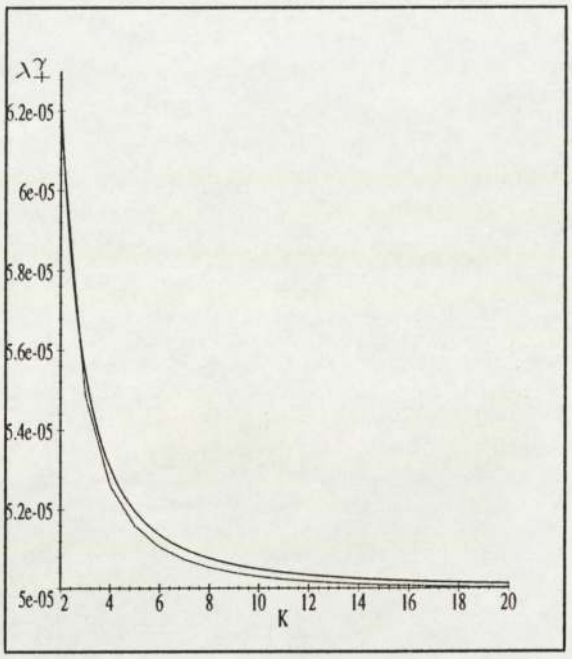


Figure 4.6: Numerical and analytical values for λ_+^γ. The solid curve corresponds to the analytical solution and the dashed curve is the numerical one. The coefficients are as follows: $\eta = 10^{-2}$ and $\gamma = 10^{-4}$.

4.3 Summary

The modified back-propagation proposed in equation 4.1 shortens the symmetric plateau significantly. However, this result is shown here for a realisable scenario only ($K = M$) and under certain assumptions for the coefficients used, nevertheless, we expect a similar behaviour for other training scenarios.

Chapter 5

Conclusion

The aim of this study was to examine the effect of task non-stationarity on the learning dynamics in an on-line learning scenario.

Non-stationarity has been modelled by rotations. This implies that the general case (any K and M) is built by 2×2 blocks representing rotations in a system with two hidden nodes. By solving numerically the dynamics for the case $K = M = 2$ it has been found that the length of the symmetric plateau is much shorter for non-stationary tasks. The theoretical study based on expanding around the fixed points and finding the positive eigenvalue of escaping the symmetric phase, confirms the numerical solutions as the positive eigenvalue is bigger than that for the stationary case.

This behaviour has given insight for suggesting a modified back-propagation algorithm speeding up the escape from the symmetric plateau. It is based on subtracting from each student vector a part of the others, speeding up their separation. This modified gradient descent is used only in the symmetric phase and has been studied for a general scenario where the number of hidden nodes of the teacher and the student are both the same. Numerical solutions show a decrease in the length of the symmetric plateau and the analytical work confirms this behaviour.

The second important phase of the learning process is the convergence phase. Unfortunately, even if the system escapes quickly from the symmetric subspace it is not converging well as the generalisation error is larger than zero for the case of non-

stationary tasks. The asymptotic values of the order parameters are found analytically and are consistent with those obtained numerically. The generalisation error is then plotted in terms of the learning rate η and the rotation frequency ω showing a divergence for certain values of η defining η_{max}.

The dependence of the residual generalisation error on η for fixed ω is curved and has a minimum for a certain value of η. This value of η enables us to have the minimal value for the generalisation error for a given rotation rate, although it has no relation with the time needed to reach this asymptotic value.

To minimise the convergence time, η^2 terms are kept in the dynamics and the dependence on η of the two largest eigenvalues of the linear system of equations around the asymptotic fixed point is studied. It enables us to find the learning rate minimising the time needed for convergence as well as the maximal learning rate.

The dependence of the learning rate optimising the time needed to reach the asymptotic point in the $K = M = 2$ case on ω is not investigated in this work due to lack of time. It also would be interesting to look at the general case for which there is any number of hidden nodes for the teacher and student and which is characterised by simultaneous rotations in several directions. This study is complicated because of the number of variables used and the new dynamics are not clear as all the vectors are rotating with different velocities.

References

[1] C.M. Bishop, *Neural networks for pattern recognition* (Oxford University Press, Oxford, 1995)

[2] G. Cybenko, *Math. Control Signals and Systems*, **2**, 303 (1989)

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

[4] P. Riegler and M. Biehl, *J. Phys. A* **28**, L506-L513 (1995)

[5] A.H.L. West and D. Saad, *Neural Information Processing System* **8**, eds. D.S Touretzky, M.C. Mozer, and M.E Hasselmo (MIT Press, 1996) p.323-329

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

[7] F.R. Gantmacher, *The theory of Matrices*, **1**, Chelsea Publishing Company New York (1960), p271-272

[8] Seth Warner, *Modern Algebra*, Dover Publication Inc., New York (1990), P748

[9] M. Biehl and H. Schwarze, *Learning drifting concepts with neural networks*, *J. Phys. A* **26** 2651 (1993)

Appendix A

Equations for a fixed environment

These results are taken from [3]. The set of dynamics is as follows:

$$\frac{dR_{in}}{d\alpha} = \eta \langle \delta_i \, y_n \rangle ,$$

$$\frac{dQ_{ik}}{d\alpha} = \eta \langle \delta_i \, x_k \rangle + \eta \langle \delta_k \, x_i \rangle + \eta^2 \langle \delta_i \, \delta_k \rangle . \tag{A.1}$$

The averages in Eq. (A.1) require the evaluation of two types of multivariate Gaussian integrals. Terms proportional to η involve the three-dimensional integral

$$I_3 \equiv \langle g'(u) \, v \, g(w) \rangle ,$$

where the argument u of g' is one of the components of $\mathbf{x}$, while both v and w can be components of either $\mathbf{x}$ or $\mathbf{y}$. The term proportional to η^2 involves the four-dimensional Gaussian integral

$$I_4 \equiv \langle g'(u) \, g'(v) \, g(w) \, g(z) \rangle ,$$

where u and v are components of $\mathbf{x}$ while w and z can be components of either $\mathbf{x}$ or $\mathbf{y}$.

The expressions for the derivatives become:

$$\begin{aligned}
\frac{dR_{in}}{d\alpha} &= \eta \left\{ \sum_m I_3(i,n,m) - \sum_j I_3(i,n,j) \right\} , \\
\frac{dQ_{ik}}{d\alpha} &= \eta \left\{ \sum_m I_3(i,k,m) - \sum_j I_3(i,k,j) \right\} + \eta \left\{ \sum_m I_3(k,i,m) - \sum_j I_3(k,i,j) \right\} + \\
&\quad \eta^2 \left\{ \sum_{n,m} I_4(i,k,n,m) - 2 \sum_{j,n} I_4(i,k,j,n) + \sum_{j,l} I_4(i,k,j,l) \right\} .
\end{aligned} \tag{A.2}$$

The arguments assigned to I_3 and I_4 correspond to the convention used to distinguish between student and teacher activation (i, j ,... for students and n, m, ... for teachers). So, $I_3(i,n,j) \equiv \langle g'(x_i)\ y_n\ g(x_j)\rangle$, and the average is performed using the three-dimensional covariance matrix C_3 which results from projecting the full covariance matrix C (defined in chapter 2) onto the relevant subspace. For $I_3(i,n,j)$ the corresponding matrix is:

$$C_3 = \begin{pmatrix} Q_{ii} & R_{in} & Q_{ij} \\ R_{in} & T_{nn} & R_{jn} \\ Q_{ij} & R_{jn} & Q_{jj} \end{pmatrix} .$$

The two multivariate integrals in Eq. (A.2) can be performed analytically for $g(x) = \mathrm{erf}(x/\sqrt{2})$. I_3 is given in terms of the components of the C_3 covariance matrix by

$$I_3 = \frac{2}{\pi}\ \frac{1}{\sqrt{\Lambda_3}}\ \frac{C_{23}(1+C_{11}) - C_{12}C_{13}}{1+C_{11}}\ , \tag{A.3}$$

with

$$\Lambda_3 = (1+C_{11})(1+C_{33}) - C_{13}^2\ . \tag{A.4}$$

The expression for I_4 in terms of the components of the corresponding C_4 covariance matrix is

$$I_4 = \frac{4}{\pi^2}\ \frac{1}{\sqrt{\Lambda_4}}\ \arcsin\left(\frac{\Lambda_0}{\sqrt{\Lambda_1}\sqrt{\Lambda_2}}\right), \tag{A.5}$$

where

$$\Lambda_4 = (1+C_{11})(1+C_{22}) - C_{12}^2\ , \tag{A.6}$$

and

$$\begin{aligned}
\Lambda_0 &= \Lambda_4 C_{34} - C_{23}C_{24}(1+C_{11}) - C_{13}C_{14}(1+C_{22}) + C_{12}C_{13}C_{24} + C_{12}C_{14}C_{23}\ , \\
\Lambda_1 &= \Lambda_4(1+C_{33}) - C_{23}^2(1+C_{11}) - C_{13}^2(1+C_{22}) + 2C_{12}C_{13}C_{23}\ , \\
\Lambda_2 &= \Lambda_4(1+C_{44}) - C_{24}^2(1+C_{11}) - C_{14}^2(1+C_{22}) + 2C_{12}C_{14}C_{24}\ .
\end{aligned} \tag{A.7}$$

Appendix B

Canonical form for an orthogonal linear transformation

Theorem:

For a real vector space $\mathcal{V}$ of finite dimension, a linear transformation ψ such that ψ: $\mathcal{V} \rightarrow \mathcal{V}$ and ψ^* the adjoint of ψ, it follows that

$\psi\psi^* = I \Leftrightarrow \exists$ an orthonormal basis $\mathcal{B}$ such that ψ is expressed as follows:

$$\psi = \begin{pmatrix} 1 & 0 & \cdots & & & & 0 \\ 0 & \ddots & & & & & \\ & & 1 & & & & \\ & & & -1 & & 0 & \\ \vdots & & & & \ddots & & \vdots \\ & & & & & -1 & \\ & & 0 & & & \boxed{A_1} & \\ & & & & & & \ddots \quad 0 \\ 0 & & & \cdots & & & \boxed{A_p} \end{pmatrix} \quad \text{and} \quad \mathbf{A}_i = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}$$

Proof:

There is an equivalence to show. The part ($\Leftarrow$) of the equivalence is easily shown by calculating $\psi\psi^*$, which is equal to I. The proof will now focus on ($\Rightarrow$) part of the equivalence.

$\forall \mathbf{x} \in \mathcal{V} \quad ||\psi(\mathbf{x})|| = ||\mathbf{x}||$ because ψ is orthogonal, therefore if $\lambda \in Sp(\psi)$[1] then $|\lambda| = 1$. But λ is either real or complex, let's then suppose that there exists a real eigenvalue denoted λ. It can have two different values: $\lambda = 1$ or $\lambda = -1$. We consider the two vector spaces: $\mathcal{E}_1 = ker(\psi - I)$ and $\mathcal{E}_{-1} = ker(\psi + I)$. They have in common the vector $\mathbf{0}$ only, they are then in a direct sum which will be denoted $\mathcal{F}$: $\mathcal{F} = \mathcal{E}_1 \oplus \mathcal{E}_{-1}$.

Lemma:

For any orthogonal operator u, if a vector space E is invariant under u then its orthogonal complement is also invariant under u.

Let's show this lemma. First of all, if $u(E) \subseteq E$ then let $\mathbf{x} \in E$ and $\mathbf{y} \in E^\perp$. We know that $u(\mathbf{x}) \in E$ so $u(\mathbf{x}) \,.\, \mathbf{y} = 0$. This is equivalent to $\mathbf{x} \,.\, u^*(\mathbf{y}) = 0$, which is valid for any $\mathbf{x}$. Therefore $u^*(\mathbf{y}) \in E^\perp$. We have shown that: $u(E) \subseteq E \Rightarrow u^*(E^\perp) \subseteq E^\perp$.

We also know that there exists a polynomial P such as $u = P(u^*)$ (basically because u and u^* have the same eigenvectors corresponding to conjugate eigenvalues), thus it is correct to write: $u(E^\perp) \subseteq E^\perp$. We have thus shown that: $u(E) \subseteq E \Rightarrow u(E^\perp) \subseteq E^\perp$. For more details on this proof, see [7].

We can notice that $\psi(\mathcal{F}) \subset \mathcal{F}$ so $\psi(\mathcal{F}^\perp) \subset \mathcal{F}^\perp$ and $\psi^*(\mathcal{F}^\perp) \subset \mathcal{F}^\perp$. The mapping $v = \psi + \psi^*$ has the property that $v^* = v$ and as v is a real mapping it can then be diagonalized [2]. The mapping v satisfies the relation $v(\mathcal{F}^\perp) \subset \mathcal{F}^\perp$ because ψ and ψ^* satisfy it, therefore it is possible to say that the mapping $v_{/\mathcal{F}^\perp}$ (restriction of v to $\mathcal{F}^\perp$)

[1] spectrum of ψ or set of its eigenvalues

[2] for the proof, see [8]

can be diagonalized. It is therefore possible to write $\exists \mathbf{x} \in \mathcal{F}^{\perp}$ $(\mathbf{x} \neq \mathbf{0})$ $\exists \lambda \in \Re$ (space of the real numbers) such that $v(\mathbf{x}) = \lambda\ \mathbf{x}$.

Then $(\psi + \psi^*)(\mathbf{x}) = \lambda\ \mathbf{x}$, therefore $\psi^2(\mathbf{x}) = -\mathbf{x}\ + \lambda\ \psi(\mathbf{x})$. So the vector space $F_1 = vect(\mathbf{x}, \psi(\mathbf{x}))$ is invariant under ψ. If $dim(F_1) = 1$ then $\exists \mu$ such that $\psi(\mathbf{x}) = \mu\ \mathbf{x}$, μ is then equal to ± 1 (as ψ is orthogonal) which implies that $\mathbf{x} \in \mathcal{F}$ but we know that $\mathbf{x} \in \mathcal{F}^{\perp}$ so $dim(F_1) = 2$. Consider the operator $\psi_1 = \psi_{/F_1}$; its determinant is ± 1 and if it is -1 then $\exists \mathbf{y} \in F_1 \cap \mathcal{F}$ which is impossible. So, ψ_1 is a rotation. Then the vector space: $\mathcal{F}^1 = E_1 \oplus E_{-1} \oplus F_1$ is invariant under ψ and we can continue the process by induction on the dimension of the vector space. We will then be able to write the whole space as a direct sum of E_1, E_2 , F_1, ..., F_m (the spaces F_i are found exactly as F_1). The application can then be written as shown in the theorem in the basis made by those of E_1, E_2 and F_i.

Appendix C

Full set of the dynamics for the $K = M = 2$ case

$$
\begin{aligned}
\frac{dR_{11}}{dt} &= 2\frac{\eta}{\pi(1+Q_{11})}\left[\frac{1+Q_{11}-R_{11}^2}{\sqrt{2+2Q_{11}-R_{11}^2}} - \frac{R_{11}R_{12}}{\sqrt{2+2Q_{11}-R_{12}^2}} - \frac{R_{11}}{\sqrt{1+2Q_{11}}}\right. \\
&- \left.\frac{R_{12}(1+Q_{11})-R_{11}Q_{12}}{\sqrt{(1+Q_{11})(1+Q_{22})-Q_{12}^2}}\right] - \omega\ R_{12} \\
\frac{dR_{12}}{dt} &= 2\frac{\eta}{\pi(1+Q_{11})}\left[\frac{1+Q_{11}-R_{12}^2}{\sqrt{2+2Q_{11}-R_{12}^2}} - \frac{R_{11}R_{12}}{\sqrt{2+2Q_{11}-R_{11}^2}} - \frac{R_{12}}{\sqrt{1+2Q_{11}}}\right. \\
&- \left.\frac{R_{22}(1+Q_{11})-R_{12}Q_{12}}{\sqrt{(1+Q_{11})(1+Q_{22})-Q_{12}^2}}\right] + \omega\ R_{11} \\
\frac{dR_{21}}{dt} &= 2\frac{\eta}{\pi(1+Q_{11})}\left[\frac{1+Q_{22}-R_{21}^2}{\sqrt{2+2Q_{22}-R_{21}^2}} - \frac{R_{21}R_{22}}{\sqrt{2+2Q_{22}-R_{22}^2}} - \frac{R_{21}}{\sqrt{1+2Q_{22}}},\right. \\
&- \left.\frac{R_{11}(1+Q_{22})-R_{21}Q_{12}}{\sqrt{(1+Q_{11})(1+Q_{22})-Q_{12}^2}}\pi(1+Q_{22})\right] - \omega\ R_{22} \\
\\
\frac{dR_{22}}{dt} &= 2\frac{\eta}{\pi(1+Q_{11})}\left[\frac{1+Q_{22}-R_{22}^2}{\sqrt{2+2Q_{22}-R_{22}^2}} - \frac{R_{21}R_{22}}{\sqrt{2+2Q_{22}-R_{21}^2}} - \frac{R_{22}}{\sqrt{1+2Q_{22}}},\right. \\
&- \left.\frac{R_{12}(1+Q_{22})-R_{22}Q_{12}}{\sqrt{(1+Q_{11})(1+Q_{22})-Q_{12}^2}}\pi(1+Q_{22})\right] + \omega\ R_{21}
\end{aligned}
$$

$$
\begin{aligned}
\frac{dQ_{11}}{dt} &= 4\frac{\eta}{\pi\ 1+Q_{11}}\left[\frac{R_{11}}{\sqrt{2+2Q_{11}-R_{11}^2}}+\frac{R_{12}}{\sqrt{2+2Q_{11}-R_{12}^2}}-\frac{Q_{11}}{\sqrt{1+2Q_{11}}}\right.\\
&\quad\left.-\frac{Q_{12}}{\sqrt{(1+Q_{11})(1+Q_{22})-Q_{12}^2}}\right]\\
&\quad+\frac{4\ \eta^2}{\pi^2\ \sqrt{1+2Q_{11}}}\left[\arcsin\left(\frac{1+2Q_{11}-2R_{11}^2}{2+4Q_{11}-2R_{12}^2}\right)+\arcsin\left(\frac{1+2Q_{11}-2R_{12}^2}{2+4Q_{11}-2R_{11}^2}\right)\right.\\
&\quad+2\arcsin\left(\frac{-2R_{11}R_{12}}{\sqrt{2+4Q_{11}-2R_{11}^2}\sqrt{2+4Q_{11}-2R_{12}^2}}\right)\\
&\quad-2\arcsin\left(\frac{R_{11}}{\sqrt{1+3Q_{11}}\sqrt{2+4Q_{11}-2R_{11}^2}}\right)\\
&\quad-2\arcsin\left(\frac{R_{22}+2Q_{11}R_{22}-2R_{12}Q_{12}}{\sqrt{1+Q_{22}+2Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}\sqrt{2+4Q_{11}-2R_{12}^2}}\right)\\
&\quad-2\arcsin\left(\frac{R_{12}}{\sqrt{1+3Q_{11}}\sqrt{2+4Q_{11}-2R_{12}^2}}\right)\\
&\quad-2\arcsin\left(\frac{(1+2Q_{11})R_{21}-2R_{11}Q_{12}}{\sqrt{1+Q_{22}+2Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}\sqrt{2+4Q_{11}-2R_{11}^2}}\right)\\
&\quad+\arcsin\left(\frac{Q_{11}}{1+3Q_{11}}\right)+\arcsin\left(\frac{Q_{22}+2Q_{11}Q_{12}-2Q_{12}^2}{1+Q_{22}+2Q_{11}+2Q_{11}Q_{12}-2Q_{12}^2}\right)\\
&\quad\left.+2\ \arcsin\left(\frac{Q_{12}}{\sqrt{1+3Q_{11}}\sqrt{1+Q_{22}+2Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
\frac{dQ_{12}}{dt} &= 2\frac{\eta}{\pi(1+Q_{22})}\left[\frac{R_{11}(1+Q_{22})-Q_{12}R_{21}}{\sqrt{2+2Q_{22}-R_{21}^2}}+\frac{R_{12}(1+Q_{22})-Q_{12}R_{22}}{\sqrt{2+2Q_{22}-R_{22}^2}}\right.\\
&\left.-\frac{Q_{11}(1+Q_{22})-Q_{12}^2}{\sqrt{1+Q_{22}+Q_{11}+Q_{11}Q_{22}-Q_{12}^2}}-\frac{Q_{12}}{\sqrt{1+2Q_{22}}}\right]\\
&+ 2\frac{\eta}{\pi(1+Q_{11})}\left[\frac{R_{21}(1+Q_{11})-Q_{12}R_{11}}{\sqrt{2+2Q_{11}-R_{11}^2}}+\frac{R_{12}(1+Q_{11})-Q_{12}R_{12}}{\sqrt{2+2Q_{11}-R_{12}^2}}\right.\\
&\left.-\frac{Q_{22}(1+Q_{11})-Q_{12}^2}{\sqrt{1+Q_{22}+Q_{11}+Q_{11}Q_{22}-Q_{12}^2}}-\frac{Q_{12}}{\sqrt{1+2Q_{11}}}\right]\\
&+ 4\frac{\eta^2}{\pi^2\sqrt{(1+Q_{22})(1+Q_{11})-Q_{12}^2}}\\
&\left[\arcsin\left(\frac{(1+Q_{11})(1+Q_{22})-Q_{12}^2-(1+Q_{11})R_{21}^2-(1+Q_{22})R_{11}^2+2Q_{12}R_{11}R_{21}}{2(1+Q_{11})(1+Q_{22})-2Q_{12}^2-R_{21}^2(1+Q_{11})-R_{11}^2(1+Q_{22})+2Q_{12}R_{11}R_{21}}\right)\right.\\
&+ \arcsin\left(\frac{(1+Q_{11})(1+Q_{22})-Q_{12}^2-(1+Q_{11})R_{22}^2-(1+Q_{22})R_{12}^2+2Q_{12}R_{12}R_{22}}{2(1+Q_{11})(1+Q_{22})-2Q_{12}^2-R_{22}^2(1+Q_{11})-R_{12}^2(1+Q_{22})+2Q_{12}R_{12}R_{22}}\right)\\
&+ 2\arcsin\left(\frac{-(1+Q_{11})R_{21}R_{22}-(1+Q_{22})R_{11}R_{12}+Q_{12}R_{11}R_{22}+Q_{12}R_{12}R_{21}}{\sqrt{2(1+Q_{11})(1+Q_{22})-2Q_{12}^2-R_{21}^2(1+Q_{11})-R_{11}^2(1+Q_{22})+2Q_{12}R_{11}R_{21}}}\right.\\
&\left.\left.\frac{1}{\sqrt{2(1+Q_{11})(1+Q_{22})-2Q_{12}^2-R_{22}^2(1+Q_{11})-R_{12}^2(1+Q_{22})+2Q_{12}R_{12}R_{22}}}\right)\right.\\
&-2\arcsin\left(\frac{(1+Q_{11})(1+Q_{22})R_{11}-Q_{12}R_{21}-Q_{11}R_{11}-Q_{22}Q_{11}R_{11}}{\sqrt{1+Q_{22}+2Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}}\right.\\
&\left.\left.\frac{1}{\sqrt{2(1+Q_{11})(1+Q_{22})-2Q_{12}^2-R_{21}^2(1+Q_{11})-R_{11}^2(1+Q_{22})+2Q_{12}R_{11}R_{21}}}\right)\right.\\
&-2\arcsin\left(\frac{(1+Q_{11})(1+Q_{22})R_{22}-Q_{22}R_{22}-Q_{12}R_{12}-Q_{11}R_{22}Q_{22}}{\sqrt{1+2Q_{22}+Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}}\right.\\
&\left.\left.\frac{1}{\sqrt{2(1+Q_{11})(1+Q_{22})-2Q_{12}^2-R_{22}^2(1+Q_{11})-R_{12}^2(1+Q_{22})+2Q_{12}R_{12}R_{22}}}\right)\right.\\
&-2\arcsin\left(\frac{(1+Q_{11})(1+Q_{22})R_{12}-Q_{12}R_{22}-Q_{11}R_{12}-Q_{11}R_{12}Q_{22}}{\sqrt{1+Q_{22}+2Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}}\right.\\
&\left.\left.\frac{1}{\sqrt{2(1+Q_{11})(1+Q_{22})-2Q_{12}^2-R_{22}^2(1+Q_{11})-R_{12}^2(1+Q_{22})+2Q_{12}R_{12}R_{22}}}\right)\right.
\end{aligned}
$$

$$
\begin{aligned}
&-2 \quad \arcsin\left(\frac{(1+Q_{11})(1+Q_{22})R_{21} - Q_{22}R_{21} - Q_{12}R_{11} - Q_{11}R_{21}Q_{22}}{\sqrt{1+2Q_{22}+Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}}\right.\\
&\qquad \left.\frac{1}{\sqrt{2(1+Q_{11})(1+Q_{22}) - 2Q_{12}^2 - R_{21}^2(1+Q_{11}) - R_{11}^2(1+Q_{22}) + 2Q_{12}R_{11}R_{21}}}\right)\\
&+ \quad \arcsin\left(\frac{(1+Q_{11})(1+Q_{22})Q_{11} - Q_{12}^2 - Q_{11}^2(1+Q_{22})}{1+Q_{22}+2Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}\right)\\
&+ \quad \arcsin\left(\frac{(1+Q_{11})(1+Q_{22})Q_{22} - Q_{22}^2(1+Q_{11}) - Q_{12}^2}{1+2Q_{22}+Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}\right)\\
&+ \quad 2\arcsin\left(\frac{(1+Q_{11})(1+Q_{22})Q_{12} - (1+Q_{11})Q_{12}Q_{22} - Q_{11}Q_{12}}{\sqrt{1+Q_{22}+2Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}\sqrt{1+2Q_{22}+Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}}\right)\Bigg].
\end{aligned}
$$

$$
\begin{aligned}
\frac{dQ_{22}}{dt} &= 4\frac{\eta}{\pi\,(1+Q_{22})}\left[\frac{R_{21}}{\sqrt{2+2Q_{22}-R_{21}^2}} + \frac{R_{22}}{\sqrt{2+2Q_{22}-R_{22}^2}} - \frac{Q_{22}}{\sqrt{1+2Q_{22}}}\right.\\
&- \quad \left.\frac{Q_{12}}{\sqrt{(1+Q_{11})(1+Q_{22})-Q_{12}^2}}\right]\\
&+ \quad \frac{4\,\eta^2}{\pi^2\,\sqrt{1+2Q_{22}}}\left[\arcsin\left(\frac{1+2Q_{22}-2R_{21}^2}{2+4Q_{22}-2R_{21}^2}\right) + \arcsin\left(\frac{1+2Q_{22}-2R_{22}^2}{2+4Q_{22}-2R_{22}^2}\right)\right.\\
&+ \quad 2\arcsin\left(\frac{-2R_{21}R_{22}}{\sqrt{2+4Q_{22}-2R_{21}^2}\sqrt{2+4Q_{22}-2R_{21}^2}}\right)\\
&- \quad 2\arcsin\left(\frac{R_{22}}{\sqrt{1+3Q_{22}}\sqrt{2+4Q_{22}-2R_{22}^2}}\right)\\
&- \quad 2\arcsin\left(\frac{R_{11}(1+2Q_{22})-2R_{21}Q_{12}}{\sqrt{1+2Q_{22}+Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}\sqrt{2+4Q_{22}-2R_{22}^2}}\right)\\
&- \quad 2\arcsin\left(\frac{R_{21}}{\sqrt{1+3Q_{22}}\sqrt{2+4Q_{11}-2R_{21}^2}}\right)\\
&- \quad 2\arcsin\left(\frac{(1+2Q_{22})R_{12}-2R_{22}Q_{12}}{\sqrt{1+Q_{22}+2Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}\sqrt{2+4Q_{22}-2R_{22}^2}}\right)\\
&+ \quad \arcsin\left(\frac{Q_{22}}{1+3Q_{22}}\right) + \arcsin\left(\frac{Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}{1+Q_{22}+2Q_{11}+2Q_{11}Q_{12}-2Q_{12}^2}\right)\\
&+ \quad 2\ \arcsin\left(\frac{Q_{12}}{\sqrt{1+3Q_{22}}\sqrt{1+Q_{22}+2Q_{11}+2Q_{11}Q_{22}-2Q_{12}^2}}\right)\Bigg]
\end{aligned}
$$

Appendix D

Asymptotic fixed point for the $K = M = 2$ case

$$
\begin{aligned}
Q_{11} &= 1 - \frac{12}{49}\,\omega\,\pi\,\big(-1890\sqrt{3}\,\pi^2\,\eta^2\,\sqrt{2} + 3465\,\eta^2\,\pi^2 - 1008\,\pi\,\eta^3\,\sqrt{3}\,\sqrt{5} - 1134\,\eta^4\,\sqrt{5} \\
&+ 2562\,\eta^4 - 1890\,\eta^3\,\sqrt{2}\,\pi + 1134\,\eta^3\,\sqrt{5}\,\pi\,\sqrt{2} + 945\,\pi^2\,\eta^2\,\sqrt{5} - 315\,\sqrt{3}\,\pi^3\,\eta \\
&- 18300\eta^3\,\omega\,\pi + 14445\,\sqrt{3}\,\pi^3\,\omega\,\eta\,\sqrt{2} - 22935\,\eta\,\pi^3\,\omega + 1785\,\sqrt{3}\,\pi^4\,\omega \\
&- 8667\,\eta^2\,\sqrt{5}\,\omega\,\pi^2\,\sqrt{2} + 8100\,\omega\,\pi\,\eta^3\,\sqrt{5} + 1554\,\sqrt{3}\,\pi\,\eta^3 + 14445\,\eta^2\,\sqrt{2}\,\omega\,\pi^2 \\
&- 8396\,\eta^2\,\sqrt{3}\,\omega\,\pi^2 + 6096\,\eta^2\,\sqrt{5}\,\sqrt{3}\,\pi^2\,\omega - 8415\,\eta\,\sqrt{5}\,\omega\,\pi^3\big)\,/ \\
&\eta^2\,\big(90\,\sqrt{3}\,\pi^2\,\eta\,\sqrt{2} - 165\,\eta\,\pi^2 + 48\,\sqrt{3}\,\pi\,\eta^2\,\sqrt{5} + 54\,\eta^3\,\sqrt{5} - 122\,\eta^3 + 90\,\pi\,\eta^2\,\sqrt{2} \\
&-54\,\pi\,\eta^2\,\sqrt{2}\,\sqrt{5} - 45\,\pi^2\,\eta\,\sqrt{5} - 74\,\eta^2\,\sqrt{3}\,\pi + 15\,\sqrt{3}\,\pi^3\big) \\
Q_{12} &= \frac{36}{49}\,\omega^2\,\pi^2\,\big(4530\,\pi^2\eta\,\sqrt{2} - 3100\,\eta^2\,\pi - 906\,\eta^2\sqrt{5}\,\pi\sqrt{6} + 864\,\eta^3\sqrt{15} \\
&- 1952\,\eta^3\sqrt{3} - 780\sqrt{3}\,\pi^2\eta\sqrt{5} - 2550\sqrt{3}\pi^2\eta + 2112\,\pi\,\eta^2\,\sqrt{5} \\
&+ 1510\,\eta^2\,\sqrt{2}\,\sqrt{3}\,\pi + 525\,\pi^3\big)\,/ \\
&\eta^2\,\big(90\,\sqrt{3}\,\pi^2\,\eta\,\sqrt{2} - 165\,\eta\,\pi^2 + 48\,\sqrt{3}\,\pi\,\eta^2\,\sqrt{5} + 54\,\eta^3\,\sqrt{5} - 122\,\eta^3 + 90\,\pi\,\eta^2\,\sqrt{2} \\
&-54\,\pi\,\eta^2\,\sqrt{2}\,\sqrt{5} - 45\,\pi^2\,\eta\,\sqrt{5} - 74\,\eta^2\,\sqrt{3}\,\pi + 15\,\sqrt{3}\,\pi^3\big)
\end{aligned}
$$

$$
\begin{aligned}
Q_{22} &= 1 + \frac{12}{49}\,\omega\,\pi\left(-1890\sqrt{3}\,\pi^2\,\eta^2\,\sqrt{2} + 3465\,\eta^2\,\pi^2 - 1008\,\pi\,\eta^3\,\sqrt{3}\,\sqrt{5} - 1134\,\eta^4\,\sqrt{5}\right.\\
&+ 2562\,\eta^4 - 1890\,\eta^3\,\sqrt{2}\,\pi + 1134\,\eta^3\,\sqrt{5}\,\pi\,\sqrt{2} + 945\,\pi^2\,\eta^2\,\sqrt{5} - 315\,\sqrt{3}\,\pi^3\,\eta\\
&- 18300\eta^3\,\omega\,\pi + 14445\,\sqrt{3}\,\pi^3\,\omega\,\eta\,\sqrt{2} - 22935\,\eta\,\pi^3\,\omega + 1785\,\sqrt{3}\,\pi^4\,\omega\\
&- 8667\,\eta^2\,\sqrt{5}\,\omega\,\pi^2\,\sqrt{2} + 8100\,\omega\,\pi\,\eta^3\,\sqrt{5} + 1554\,\sqrt{3}\,\pi\,\eta^3 + 14445\,\eta^2\,\sqrt{2}\,\omega\,\pi^2\\
&\left.- 8396\,\eta^2\,\sqrt{3}\,\omega\,\pi^2 + 6096\,\eta^2\,\sqrt{5}\,\sqrt{3}\,\pi^2\,\omega - 8415\,\eta\,\sqrt{5}\,\omega\,\pi^3\right)/\\
&\eta^2\left(90\,\sqrt{3}\,\pi^2\,\eta\,\sqrt{2} - 165\,\eta\,\pi^2 + 48\,\sqrt{3}\,\pi\,\eta^2\,\sqrt{5} + 54\,\eta^3\,\sqrt{5} - 122\,\eta^3 + 90\,\pi\,\eta^2\,\sqrt{2}\right.\\
&\left.-54\,\pi\,\eta^2\,\sqrt{2}\,\sqrt{5} - 45\,\pi^2\,\eta\,\sqrt{5} - 74\,\eta^2\,\sqrt{3}\,\pi + 15\,\sqrt{3}\,\pi^3\right)\\
R_{11} &= 1 - \frac{9}{49}\,\omega\,\pi\left(-1260\,\sqrt{3}\,\pi^2\,\eta^2\,\sqrt{2} + 2310\eta^2\,\pi^2 - 672\,\pi\,\eta^3\sqrt{3}\,\sqrt{5} - 756\,\eta^4\,\sqrt{5}\right.\\
&+ 1708\,\eta^4 - 1260\,\eta^3\sqrt{2}\pi + 756\eta^3\sqrt{5}\pi\sqrt{2} + 630\pi^2\eta^2\sqrt{5} + 1036\eta^3\,\sqrt{3}\pi\\
&- 210\sqrt{3}\pi^3\,\eta - 12200\eta^3\,\omega\pi + 9315\sqrt{3}\,\pi^3\,\omega\,\pi^2 + 4176\eta^2\sqrt{5}\sqrt{3}\pi^2\,\omega\\
&+ 1785\,\sqrt{3}\pi^4\omega + 5400\omega\pi\eta^3\sqrt{5} - 15255\,\eta\,\pi^3\omega - 5784\eta^2\sqrt{3}\,\omega\pi^2\\
&\left.+ 9315\,\eta^2\sqrt{2}\omega\,\pi^2 - 5967\,\eta\,\sqrt{5}\omega\pi^3 - 5589\,\eta^2\,\sqrt{5}\,\omega\,\pi^2\,\sqrt{2}\right)/\\
&\eta^2\left(90\,\sqrt{3}\,\pi^2\,\eta\,\sqrt{2} - 165\,\eta\,\pi^2 + 48\,\sqrt{3}\,\pi\,\eta^2\,\sqrt{5} + 54\,\eta^3\,\sqrt{5} - 122\,\eta^3 + 90\,\pi\,\eta^2\,\sqrt{2}\right.\\
&\left.-54\,\pi\,\eta^2\,\sqrt{2}\,\sqrt{5} - 45\,\pi^2\,\eta\,\sqrt{5} - 74\,\eta^2\,\sqrt{3}\,\pi + 15\,\sqrt{3}\,\pi^3\right)\\
R_{12} &= \frac{4}{49}\,\omega\pi\left(-3108\,\eta^3\pi - 1708\eta^4\sqrt{3} + 630\eta\pi^3 + 756\eta^4\sqrt{3}\sqrt{5} + 3780\pi^2\,\eta^2\sqrt{2}\right.\\
&- 630\sqrt{3}\pi^2\eta^2\sqrt{5} - 756\,\eta^3\,\sqrt{5}\,\pi\,\sqrt{6} + 2016\,\eta^3\,\pi\,\sqrt{5} - 2310\,\eta^2\sqrt{3}\pi^2\\
&+ 1260\,\eta^3\sqrt{6}\pi + 6480\,\eta^2\sqrt{6}\,\omega\,\pi^2 - 8784\,\eta^3\,\sqrt{3}\,\omega\pi - 11160\eta\sqrt{3}\pi^3\omega\\
&- 13572\,\eta^2\,\omega\,\pi^2 + 9504\,\omega\,\pi^2\,\eta^2\,\sqrt{5} - 3888\,\omega\pi^2\eta^2\sqrt{30}\\
&\left.+ 4725\,\pi^4\omega + 19440\,\omega\,\pi^3\,\eta\,\sqrt{2} + 3888\,\omega\,\pi\,\eta^3\,\sqrt{15} - 4266\eta\,\pi^3\,\omega\,\sqrt{15}\right)/\\
&\eta^2\left(90\,\sqrt{3}\,\pi^2\,\eta\,\sqrt{2} - 165\,\eta\,\pi^2 + 48\,\sqrt{3}\,\pi\,\eta^2\,\sqrt{5} + 54\,\eta^3\,\sqrt{5} - 122\,\eta^3 + 90\,\pi\,\eta^2\,\sqrt{2}\right.\\
&\left.-54\,\pi\,\eta^2\,\sqrt{2}\,\sqrt{5} - 45\,\pi^2\,\eta\,\sqrt{5} - 74\,\eta^2\,\sqrt{3}\,\pi + 15\,\sqrt{3}\,\pi^3\right)
\end{aligned}
$$

$$
\begin{aligned}
R_{21} &= -\frac{4}{49}\,\omega\pi\left(-3108\,\eta^3\pi - 1708\eta^4\sqrt{3} + 630\eta\pi^3 + 756\eta^4\sqrt{3}\sqrt{5} + 3780\pi^2\,\eta^2\sqrt{2}\right.\\
&- 630\sqrt{3}\pi^2\eta^2\sqrt{5} - 756\,\eta^3\,\sqrt{5}\,\pi\,\sqrt{6} + 2016\,\eta^3\,\pi\,\sqrt{5} - 2310\,\eta^2\sqrt{3}\pi^2\\
&+ 1260\,\eta^3\sqrt{6}\pi + 6480\,\eta^2\sqrt{6}\,\omega\,\pi^2 - 8784\,\eta^3\,\sqrt{3}\,\omega\pi - 11160\eta\sqrt{3}\pi^3\omega\\
&- 13572\,\eta^2\,\omega\,\pi^2 + 9504\,\omega\,\pi^2\,\eta^2\,\sqrt{5} - 3888\,\omega\pi^2\eta^2\sqrt{30}\\
&+ \left.4725\,\pi^4\omega + 19440\,\omega\,\pi^3\,\eta\,\sqrt{2} + 3888\,\omega\,\pi\,\eta^3\,\sqrt{15} - 4266\eta\,\pi^3\,\omega\,\sqrt{15}\right)/\\
&\eta^2\left(90\,\sqrt{3}\,\pi^2\,\eta\,\sqrt{2} - 165\,\eta\,\pi^2 + 48\,\sqrt{3}\,\pi\,\eta^2\,\sqrt{5} + 54\,\eta^3\,\sqrt{5} - 122\,\eta^3 + 90\,\pi\,\eta^2\,\sqrt{2}\right.\\
&\left.-54\,\pi\,\eta^2\,\sqrt{2}\,\sqrt{5} - 45\,\pi^2\,\eta\,\sqrt{5} - 74\,\eta^2\,\sqrt{3}\,\pi + 15\,\sqrt{3}\,\pi^3\right)\\
R_{22} &= 1 + \frac{9}{49}\,\omega\,\pi\left(-1260\,\sqrt{3}\,\pi^2\,\eta^2\,\sqrt{2} + 2310\eta^2\,\pi^2 - 672\,\pi\,\eta^3\sqrt{3}\,\sqrt{5} - 756\,\eta^4\,\sqrt{5}\right.\\
&+ 1708\,\eta^4 - 1260\,\eta^3\sqrt{2}\pi + 756\eta^3\sqrt{5}\pi\sqrt{2} + 630\pi^2\eta^2\sqrt{5} + 1036\eta^3\,\sqrt{3}\pi\\
&- 210\sqrt{3}\pi^3\,\eta - 12200\eta^3\,\omega\pi + 9315\sqrt{3}\,\pi^3\,\omega\,\pi^2 + 4176\eta^2\sqrt{5}\sqrt{3}\pi^2\,\omega\\
&+ 1785\,\sqrt{3}\pi^4\omega + 5400\omega\pi\eta^3\sqrt{5} - 15255\,\eta\,\pi^3\omega - 5784\eta^2\sqrt{3}\,\omega\pi^2\\
&+ \left.9315\,\eta^2\sqrt{2}\omega\,\pi^2 - 5967\,\eta\,\sqrt{5}\omega\pi^3 - 5589\,\eta^2\,\sqrt{5}\,\omega\,\pi^2\,\sqrt{2}\right)/\\
&\eta^2\left(90\,\sqrt{3}\,\pi^2\,\eta\,\sqrt{2} - 165\,\eta\,\pi^2 + 48\,\sqrt{3}\,\pi\,\eta^2\,\sqrt{5} + 54\,\eta^3\,\sqrt{5} - 122\,\eta^3 + 90\,\pi\,\eta^2\,\sqrt{2}\right.\\
&\left.-54\,\pi\,\eta^2\,\sqrt{2}\,\sqrt{5} - 45\,\pi^2\,\eta\,\sqrt{5} - 74\,\eta^2\,\sqrt{3}\,\pi + 15\,\sqrt{3}\,\pi^3\right).
\end{aligned}
$$