Probing the basins of attraction of a recurrent neural network

2000 J. Phys. A: Math. Gen. 33 1765

(http://iopscience.iop.org/0305-4470/33/9/304)

J. Phys. A: Math. Gen. **33** (2000) 1765–1780. Printed in the UK PII: S0305-4470(00)06037-6

Probing the basins of attraction of a recurrent neural network

M Heerema and W A van Leeuwen

Institute for Theoretical Physics, University of Amsterdam, Valckenierstraat 65, 1018 XE Amsterdam, The Netherlands

E-mail: heerema@wins.uva.nl and leeuwen@wins.uva.nl

Received 13 July 1999, in final form 20 December 1999

Abstract. A recurrent neural network is considered that can retrieve a collection of patterns, as well as slightly perturbed versions of this 'pure' set of patterns via fixed points of its dynamics. By replacing the set of dynamical constraints, i.e. the fixed point equations, by an extended collection of fixed-point-like equations, analytical expressions are found for the weights $w_{ij}(b)$ of the net, which depend on a certain parameter b. This so-called basin parameter b is such that for $b = 0$ there are, *a priori*, no perturbed patterns to be recognized by the net. It is shown by a numerical study, via probing sets, that a net constructed to recognize perturbed patterns, i.e. with values of the connections $w_{ij}(b)$ with $b \neq 0$, possesses larger basins of attraction than a net made with the help of a pure set of patterns, i.e. with connections $w_{ij}(b = 0)$. The mathematical results obtained can, in principle, be realized by an actual, biological neural net.

1. Introduction

The capacity of a neural network to recognize a pattern that is not precisely equal to, but resembles, a given, stored pattern is characterized by what is called, in a mathematical context, the 'basin of attraction' of the stored pattern. If the basin is small, the network will only be capable of associating a small set of similar patterns to a typical pattern, whereas for a large basin the set of similar patterns that can be recognized is large.

Once a pattern has been presented to a neural network, the neural network starts to evolve under the influence of its own internal dynamics. If the network, at the end of this process, ends in a unique state this state is called a (single) attractor of the network. It is also possible that the network hops between more than one final state, in which case one speaks of a multiple attractor [1–3]. Patterns that evolve to an attractor are said to belong to the basin of attraction of this attractor. Many ways of characterizing basins are in vogue: basins are said to be deep or shallow and narrow or wide [4].

A way to influence the basins of the attractors is to change the network dynamics, switching from deterministic to stochastic dynamics [5, 6]. Another way to change the dynamics of the neural network is to vary the connections during the learning stage. The latter possibility can be exploited in a model for an actual, biological system [7–9].

We are primarily interested in biological neural networks. Therefore, we are not aiming at mathematical problems such as (optimizing) the storage capacity in relation to the sizes of the basins of attraction, a subject that has received ample attention in the literature [10–14].

Many dynamical systems are parametrized by a certain constant κ, sometimes called the 'margin parameter'. The margin parameter κ is claimed to be related to the size of the basins of attraction of the fixed points of the dynamics of a neural network [15–19]. Naively, one would expect, for reasons that are directly related to the way this parameter κ is introduced in the model, that the larger is κ, the larger the basins of attraction will be. However, as the 1997 study of Rodrigues Neto and Fontanari indicates, this may not be true. Their numerical analysis for tiny networks (up to 24 neurons) suggests that the number of attractors increases with increasing κ and that, perhaps because of this increase, the basins of attraction are not enlarged—as one might at first expect [20]. In section 3, we arrive at a precise interpretation for the margin parameter κ, which, usually, is introduced as an *ad hoc* quantity. In section 4, we consider the effect of the margin parameter on the basins for a network with 256 neurons. For a basin parameter $b = 0$, we find that the larger is κ, the larger the basins, in agreement with what one would expect naively (see figure 1 for $b = 0$).

The 1992 study of Wong and Sherrington is also concerned with the sizes of the basins of attraction. One of their finding is, roughly speaking, that the noisier the set of learning patterns, the larger the basins of attraction [4]. Our findings support their observations (see figure 1 for $b > 0$).

We consider a network in its final state only, i.e. after the process of learning has stopped. This makes our study time-independent. We try and construct a network with weights w_{ij} that can not only store a certain set of p prescribed patterns $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)$, where $\mu = 1, \ldots, p$, but that can also remember a larger set of patterns, centred around these typical patterns. These enlarged sets, called $\Omega^\mu(b)$ below, are characterized by the basin parameter b mentioned above. If $b = 0$, the set $\Omega^\mu(b)$ reduces to the sole pattern $\boldsymbol{\xi}^\mu$. What we obtain, finally, are values for the weights that depend on this basin parameter b:

$$w_{ij}(t) = \begin{cases} w_{ij}(t_0) + v_{ij}(t) & (j \in V_i) \\ w_{ij}(t_0) & (j \in V_i^c) \end{cases} \tag{1}$$

with

$$v_{ij}(t) = N^{-1} \sum_{\mu,\nu=1}^{p} [\kappa - \bar{\gamma}_i^\mu(b, \boldsymbol{w}_i(t_0))](2\xi_i^\mu - 1)(\bar{C}_i^{-1}(b))^{\mu\nu}[(1-b)\xi_j^\nu + b(1-\xi_j^\nu)] \tag{2}$$

(see section 5, equations (32), (34) and (37)). Here, the $w_{ij}(t_0)$ are arbitrary numbers, which can be interpreted, in a different context, as initial values for the weights, at an initial time t_0, as is suggested by the notation. Furthermore, N is the number of neurons of the network. We abbreviated $\boldsymbol{w}_i(t_0) := (w_{i1}(t_0), \ldots, w_{iN}(t_0))$. The quantities $\bar{\gamma}_i^\mu$ are defined in (15) for arbitrary $\boldsymbol{w}_i$, and the matrices $\bar{C}_i^{\mu\nu}$ are defined in (24). The $\bar{\gamma}_i^\mu$ depend on threshold potentials θ_i, the basin parameter b and the input patterns $\boldsymbol{\xi}^\mu$. V_i and V_i^c are index collections defined in such a way that the weights w_{ij} are adaptable if $j \in V_i$ and constant if $j \in V_i^c$ (for all $i = 1, \ldots, N$). If the (constants) w_{ij} vanish, there is no connection between i and j. Hence, the w_{ij} $(j \in V_i^c)$ determine the network topology. The more w_{ij} $(j \in V_i^c)$ vanish, the more 'diluted' a network. We have written $w_{ij}(t)$ for the weights, to facilitate comparison with earlier result (see, e.g., [21]). In this paper they are time independent constants, however.

The formulae (1) and (2), constituting the main result of this paper, generalize well-known results for the weights of a recurrent network. The generalizations concerned are: the network may not be fully connected, and the weights may depend upon the prescribed sizes of the basins, characterized by the basin parameter b. For $b = 0$ we recover our earlier result for a diluted network [21].

It turns out that in some cases the basins of attraction are larger for values of b unequal to zero. In other words, a network which has learned not only a set of patterns $\boldsymbol{\xi}^\mu$, but a collection of perturbed patterns $\Omega^\mu(b)$, will possess larger basins of attraction. Hence, a network can optimally recognize perturbed patterns, if it has been constructed with perturbed patterns. This is what Wong and Sherrington [4], in a related study, but for a network with connections that are changed during a learning process, call the 'principle of adaptation': a neural network is found to perform best in an operating environment identical to the training environment. Our analysis of the system after the process of learning is completed confirms this observation, albeit that the word 'identical' is not to be taken literally. So far we have given a general introduction to the problem. We now come to a short overview of our paper.

In section 2 we start by defining mathematically the problem to find suitable synaptic weights by formulating the equations to be obeyed by the weights w_{ij} of the connections. In section 3.1 and the appendix we indicate how we could obtain, in principle, a series expansion in the parameter b for the solution of the equations. To actually calculate the first terms of the expansion would be very time consuming. We therefore proceed differently. In section 3.2 we rewrite the implicit expression found in section 3.1 in such a way that we can easily find an approximation (see equation (31)). What we essentially do is to replace in the alternative implicit expression found in section 3.2, a certain average $\bar{\gamma}_i^\mu$ related to the ith neuron potential h_i, threshold θ_i and the activity x_i, given explicitly by equation (7) below, by one and the same constant κ. We thus find, by identification of κ in an old result and the κ introduced here, an interpretation for the margin parameter κ. Whether or not this replacement of the functions $\bar{\gamma}_i^\mu$ by one and the same constant makes sense is studied in the next section. In section 4 we introduce a probing set, characterized by a probing parameter $\bar{b}$. The network's performance, as a function of the basin parameter b, is calculated numerically for different values of the probing parameter $\bar{b}$. We thus test our approximation to the exact solution, and find it to be quite satisfactory.

2. Mathematical formulation of the problem

2.1. Equations for the enlarged sets of input patterns

Consider a recurrent network of N neurons. A neuron i of this network fires ($x_i = 1$) if its potential $h_i = \sum_{l=1}^N w_{il} x_l$ surpasses a certain threshold value θ_i ($i = 1, \ldots, N$). The dynamics of the network is given by the deterministic equation

$$x_i(t + \Delta t) = \Theta_\mathrm{H}\left(\sum_{l=1}^N w_{il} x_l(t) - \theta_i\right) \qquad (i = 1, \ldots, N) \tag{3}$$

where Θ_H is the Heaviside step function: $\Theta_\mathrm{H}(z) = 1$ for positive z and zero elsewhere. The weights of the neurons of the network will be updated simultaneously, i.e. we use parallel dynamics.

Let us suppose that the network is such that it can store the p patterns $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p$, where $\boldsymbol{\xi}$ is an N-dimensional vector consisting of zeros and ones. For the μth pattern we have $h_i = \sum_{l=1}^N w_{il} \xi_l^\mu$, hence the weights w_{il} of this network are constrained by the fixed point equations, following from (3)

$$\Theta_\mathrm{H}\left(\sum_{l=1}^N w_{il} \xi_l^\mu - \theta_i\right) = \xi_i^\mu \qquad (\mu = 1, \ldots, p; i = 1, \ldots, N) \tag{4}$$

(see, e.g., [21]).

Once the weights w_{il} occurring in (4) have been determined for chosen collections of patterns $\boldsymbol{\xi}^\mu$, one may ask the question which patterns $\boldsymbol{x}$, alike but not exactly equal to one of the $\boldsymbol{\xi}^\mu$, evolve to the fixed point $\boldsymbol{\xi}^\mu$, i.e. what are the basins of attraction of the fixed points $\boldsymbol{\xi}^\mu$. It is precisely the purpose of this paper to study this question in some detail.

The basin of attraction of an attractor of a dynamical system is defined to be the collection of vectors that evolve, in one or many steps, to this attractor. We are here interested in the question which vectors $\boldsymbol{x}$, belonging to certain disjunct sets of patterns Ω^μ, centred around typical patterns $\boldsymbol{\xi}^\mu$ ($\mu = 1, \ldots, p$), arrive, in one step of the dynamics only, at the fixed point $\boldsymbol{\xi}^\mu$. These latter $\boldsymbol{x}$ belong certainly to the basin of attraction as defined above, and will be referred to, for the sake of simplicity, as the basin of attraction, although it is only a part, namely, the one-step part, of the actual basin of attraction.

In order to take our newly defined basin of attraction into account, we shall replace the requirement (4), an equation for the weights w_{il}, by

$$\Theta_\mathrm{H}\left(\sum_{l=1}^{N} w_{il}x_l - \theta_i\right) = \xi_i^\mu \qquad (\mu = 1, \ldots, p; i = 1, \ldots, N) \tag{5}$$

where the patterns $\boldsymbol{x}$ belong to certain given disjunct sets of patterns Ω^μ, still to be specified, centred around typical patterns $\boldsymbol{\xi}^\mu$ ($\mu = 1, \ldots, p$). Equation (5) is the central equation of this paper; we are no longer concerned with equations (3) or (4). Note that this equation is time independent; nevertheless, we will indicate the final solution for the weights by $w_{ij}(t)$, in order to suggest that these are the weights after a period of learning. We shall determine by an (approximating) *analytical* procedure the weights w_{il} such that (5) is probably satisfied for most of the patterns $\boldsymbol{x}$, but not necessarily for all patterns. The latter will depend on the chosen collections Ω^μ ($\mu = 1, \ldots, p$). Having obtained the weights w_{il}, for such a particular choice of Ω^μ, we shall check by a *numerical* procedure whether all $\boldsymbol{x} \in \Omega^\mu$ actually satisfy (5). This will indeed not always, but often, be the case. Thus we shall have obtained values for the weights which could be useful for an actual network.

As stated above, we are not concerned, in this paper, with the process via which learning takes place, we are only studying the purely mathematical problem of finding values for the weights w_{ij} that guarantee storage and retrieval properties of a neural net. This leaves us with the question of whether the values, given by such a dry, mathematical requirement can actually be realized by the wet-ware constituted by the neurons and their connections. This point is the subject of our next paper [22], where it will turn out that a biological system can realize values for the weights which very closely approximate the values obtained here: compare formulae (1) and (2) with (38) and (39) of [22].

Distinguishing the cases $\xi_i^\mu = 0$ (no neuron activity) and $\xi_i^\mu = 1$ one may verify that the equations (5) are fulfilled if and only if

$$\gamma_i^\mu(\boldsymbol{w}_i) := \left(\sum_{l=1}^{N} w_{il}x_l - \theta_i\right)(2\xi_i^\mu - 1) > 0 \qquad \forall \boldsymbol{x} \in \Omega^\mu \tag{6}$$

where we abbreviated $\boldsymbol{w}_i := (w_{i1}, \ldots, w_{iN})$, and where Ω^μ is a collection of patterns which will be made explicit in section 2.2. Let $p^\mu(\boldsymbol{x})$ be the probability of occurrence of a pattern $\boldsymbol{x}$ in the set Ω^μ of patterns centred around a typical pattern $\boldsymbol{\xi}^\mu$. From (6) it follows that the averages $\bar{\gamma}_i^\mu$ defined as

$$\bar{\gamma}_i^\mu(\boldsymbol{w}_i) := \sum_{\boldsymbol{x}\in\Omega^\mu} p^\mu(\boldsymbol{x})\left(\sum_{l=1}^{N} w_{il}x_l - \theta_i\right)(2\xi_i^\mu - 1) \qquad (\mu = 1, \ldots, p; i = 1, \ldots, N) \tag{7}$$

are also positive, i.e.

$$\bar{\gamma}_i^\mu(\boldsymbol{w}_i) > 0. \tag{8}$$

Conversely, the fact that the averages are positive, $\bar{\gamma}_i^\mu > 0$, does not necessarily imply that $\gamma_i^\mu > 0$ ($i = 1, \ldots, N; \mu = 1, \ldots, p$). Throughout this paper, the averages $\bar{\gamma}_i^\mu$ will play a central role.

Let n_Ω be the total number of patterns belonging to any of the collections Ω^μ ($\mu = 1, \ldots, p$). Since, in general, the number n_Ω is larger than the number p, the set of equations (5) will be more restrictive than the set (4).

In the following, we shall consider biological networks, for which $w_{ii} = 0\,(i = 1, \ldots, N)$. Moreover, we shall consider partially connected (or diluted) networks, i.e. we allow for the possibility that a particular set of connections w_{ij} vanish. In general, we shall suppose that a certain subset of connections w_{ij} have prescribed values, which may or may not be zero. In order to formalize this, we introduce the sets V_i $(i = 1, \ldots, N)$ and their complements V_i^c: the V_i contain all indices j for which w_{ij} is not prescribed, but are to be determined via equation (5), while their complements V_i^c contain all indices j for which w_{ij} have certain prescribed values, which may or may not be zero [21]. Let the total number of indices j for which w_{ij} $(i = 1, \ldots, N)$ is prescribed be given by M. Then (5) is a set of Nn_Ω inequalities to be satisfied by the $N^2 - M$ unknown weights w_{ij}.

Multiplying both sides of (5) by $p^\mu(\boldsymbol{x})x_j$ and summing over μ and $\boldsymbol{x}$ we obtain

$$\sum_{\mu=1}^{p} \sum_{\boldsymbol{x}\in\Omega^\mu} p^\mu(\boldsymbol{x}) x_j \Theta_{\mathrm{H}}\left(\sum_{l=1}^{N} w_{il} x_l - \theta_i \right) = \sum_{\mu=1}^{p} \sum_{\boldsymbol{x}\in\Omega^\mu} p^\mu(\boldsymbol{x}) x_j \xi_i^\mu \qquad (j \in V_i). \tag{9}$$

These are $N^2 - M$ equations for the $N^2 - M$ non-prescribed weights w_{ij}, from which we want to solve the w_{ij}, once the Ω^μ, or, equivalently, the $p^\mu(\boldsymbol{x})$ are specified. Notwithstanding the fact that the number of equations equals the number of unknowns, the solution of (9) for the weights w_{il} is not unique, because the step function Θ_H only requires that $\sum_{l=1}^{N} w_{il}x_l - \theta_i$ be positive or negative. As an aside, note that equation (4) is underdetermined for $p < N$: then there are more unknowns w_{il} than equations.

2.2. The distribution of patterns in the basins

We choose the following, particular, probability distribution function:

$$p^\mu(\boldsymbol{x}) = \prod_{i=1}^{N} p_i^\mu(x_i) \tag{10}$$

where

$$p_i^\mu(x_i) = (1-b)\delta_{x_i,\xi_i^\mu} + b\delta_{x_i,1-\xi_i^\mu} \tag{11}$$

and where b is a parameter between 0 and 1, which we will refer to as the 'basin parameter'.

The sets Ω^μ around the patterns $\boldsymbol{\xi}^\mu$ are supposed to be disjunct, and a vector $\boldsymbol{x}$ outside $\cup_{\mu=1}^{p}\Omega^\mu$ has, by definition, a vanishing probability of occurrence. The probability distribution (10) and (11), however, yields a finite—albeit very small—probability of occurrence for a vector $\boldsymbol{x}$ outside the direct surrounding Ω^μ of $\boldsymbol{\xi}^\mu$, since it is defined for all 2^N possible vectors $\boldsymbol{x}$. The observation that the probability distribution (10) and (11) for $\boldsymbol{x}$ outside Ω^μ is very small allows us to approximate the sum of all $\boldsymbol{x} \in \cup_{\mu=1}^{p}\Omega^\mu$ by the larger sum over all $\boldsymbol{x} \in \{0, 1\}^N$. This approximation will enable us to obtain analytical results.

If $b = 0$, only the patterns $\boldsymbol{x} = \boldsymbol{\xi}^\mu$ have a non-zero probability of occurrence. For values of b close to zero any vector $\boldsymbol{x}$ has a non-zero probability of occurrence, but only vectors $\boldsymbol{x}$ close to one of the $\boldsymbol{\xi}^\mu$ have a probability of occurrence comparable to the probability of occurrence of a typical pattern. Note that the basin parameter is directly related to the magnitude of Ω^μ: the larger b, the larger the number of patterns in Ω^μ that resemble the pattern $\boldsymbol{\xi}^\mu$. Let us denote the average over the patterns as

$$\bar{x}_j^\mu := \sum_{\boldsymbol{x}\in\{0,1\}^N} p^\mu(\boldsymbol{x}) x_j. \tag{12}$$

Then, from (10) and (11) we find

$$\sum_{\boldsymbol{x}\in\{0,1\}^N} p^\mu(\boldsymbol{x}) = 1 \tag{13}$$

and

$$\bar{x}_j^\mu = \sum_{x_j=0,1} [(1-b)\delta_{x_j,\xi_j^\mu} + b\delta_{x_j,1-\xi_j^\mu}]x_j \prod_{k\neq j} \sum_{x_k=0,1} [(1-b)\delta_{x_k,\xi_k^\mu} + b\delta_{x_k,1-\xi_k^\mu}]$$
$$= (1-b)\xi_j^\mu + b(1-\xi_j^\mu). \tag{14}$$

The first equation, equation (13), expresses the normalization of the probability distribution function, the second one, equation (14), expresses the fact that the average value of the activity of neuron j is a number between 0 and 1, depending on the basin parameter b. Using (12) and (13) in (7) yields

$$\bar{\gamma}_i^\mu(b, \boldsymbol{w}_i) = \left(\sum_{l=1}^{N} w_{il}\bar{x}_l^\mu - \theta_i \right)(2\xi_i^\mu - 1) \tag{15}$$

where b is the basin parameter and where $\bar{x}_l^\mu$ is given by (14). The w_{il} occurring in this expression still have to be found.

3. Solving the equations

We will now try and solve the problem of finding the weights w_{il} of a recurrent neural network, in the approximation dictated by equation (9) combined with the particular probability distribution (10) and (11), and we hope, thereby, to have obtained a useful solution for the problem that we actually want to solve, i.e. equation (5) or, equivalently, (6) for given collections Ω^μ. The question, to what extent we will have achieved this goal will be answered in section 4, where we perform a numerical analysis.

The analytical approach to the problem to solve (9), an equation for the weights of a many-neuron recurrent network, is an adapted version of the way in which Wiegerinck and Coolen calculated the weights for a large perceptron [23].

3.1. Implicit equations for the weights

By substituting (10) and (11) into equation (9), we can obtain explicit expressions for both its left and its right side, and, from these, solve for the weights w_{il}. Using (12), we immediately obtain for the right-hand side of (9)

$$\sum_{\mu=1}^{p} \sum_{\boldsymbol{x}\in\Omega^\mu} p^\mu(\boldsymbol{x}) x_j \xi_i^\mu = \sum_\mu \xi_i^\mu \bar{x}_j^\mu \tag{16}$$

where $\bar{x}_j^\mu$ is given by (14). We turn now to the left-hand side of equation (9), the handling of which is more complicated and will be largely done in the appendix.

We note that if $w_{ij} = w_{ij}(\theta_i, \xi_i^\mu)$ is a solution of equation (4) or (5), then also $\hat{w}_{ij}(\hat{\theta}_i, \xi_i^\mu) := a_i w_{ij}(a_i^{-1}\hat{\theta}_i, \xi_i^\mu)$ is a solution of equation (4) or (5), if θ_i is replaced by $\hat{\theta}_i = a_i\theta_i$, where a_i is an arbitrary real constant. Using this freedom of gauge with $a_i = (\sum_{m=1}^N w_{im}^2)^{1/2}$, we can adjust the order of magnitude of the weights and the thresholds

$$\hat{w}_{ij} = \frac{w_{ij}}{\sqrt{\sum_{m=1}^N w_{im}^2}} \qquad \hat{\theta}_i = \frac{\theta_i}{\sqrt{\sum_{m=1}^N w_{im}^2}} \tag{17}$$

which has a consequence that, if w_{ij} and θ_i are of the order N^y (y an arbitrary real number), the hatted quantities are small, namely of the order $1/\sqrt{N}$. Note that

$$\sum_{m=1}^{N} \hat{w}_{im}^2 = 1. \tag{18}$$

Equations (17) and (18) enable us to switch, at any moment, from hatted to unhatted quantities. The hatted quantities are useful in view of the property (17), a property that is used in the appendix. One has, trivially,

$$\Theta_{\mathrm{H}}\left(\sum_{l=1}^{N} w_{il}x_l - \theta_i\right) = \Theta_{\mathrm{H}}\left(\sum_{l=1}^{N} \hat{w}_{il}x_l - \hat{\theta}_i\right). \tag{19}$$

The further evaluation of the left-hand side of (9) in terms of the $\hat{w}_{ij}$ is rather complicated and is given in the appendix. Combining the right-hand side, equation (16), and the left-hand side, equation (A.21), we find an implicit equation for the $\hat{w}_{ij}$

$$\hat{w}_{ij} = N^{-1}\sum_{\mu=1}^{p} E_i^{\mu}(b)(2\xi_i^{\mu} - 1)\bar{x}_j^{\mu} \qquad (i = 1, \ldots, N;\ j \in V_i) \tag{20}$$

where the E_i^{μ} given by

$$E_i^{\mu}(b) = N\frac{(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i))^{-1}\exp\left(-(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i))^2/2\sigma\right)}{\sum_{\mu}\exp\left(-(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i))^2/2\sigma\right)} \tag{21}$$

are positive quantities. In the latter equations we abbreviated $\sigma = b(1-b)$ and introduced $\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)$, quantities like the $\bar{\gamma}_i^{\mu}$, equation (15), of which the precise definition is given in the appendix by (A.17). With (20) and (21) we have obtained an expression for the weights $\hat{w}_{ij}$ in terms of the $\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)$, which, in turn, is a given function of the weights $\hat{w}_{ij}$, the thresholds $\hat{\theta}_i$ and the patterns $\boldsymbol{\xi}^{\mu}$. In other words, equations (20) and (21) are implicit expressions for the weights only.

We could find explicit expressions for the weights by expanding the $\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)$ as a power series in the basin parameter b

$$\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i) = \hat{\bar{\gamma}}_i^{\mu 0} + \hat{\bar{\gamma}}_i^{\mu 1}b^1 + \hat{\bar{\gamma}}_i^{\mu 2}b^2 + \cdots. \tag{22}$$

Inserting this expansion into (20) and (21), using (A.10), and equating equal powers of the expansion variable b, we may obtain explicit expressions for the expansion coefficients $\hat{\bar{\gamma}}_i^{\mu k}$ ($\mu = 1, \ldots, p; i = 1, \ldots, N; k = 0, 1, 2, \ldots, \infty$) of the power series in b, in terms of the physical quantities $\boldsymbol{\xi}^{\mu}$, $\hat{\theta}_i$ and $\hat{w}_{ij}$, where j is restricted to the set V_i^c. We thus would find an analytical solution of equation (9). This scheme has been carried out by Wiegerinck and Coolen [23] for the perceptron. We do not pursue this path for the recurrent neural net considered here, but we will use a pragmatic shortcut to arrive at an approximate explicit expression instead. This will be done on the basis of an alternative implicit expression for the weights (20), to be derived in the next section (see equation (27) below).

3.2. An alternative implicit expression for the weights

Rewriting (A.17), we may derive an alternative expression for $E_i^{\mu}(b)$. To that end we substitute (20) into (A.17):

$$\sum_{\nu=1}^{p} \bar{C}_i^{\mu\nu} E_i^{\nu}(b)(2\xi_i^{\nu} - 1) = \Gamma_i^{\mu}(b) \tag{23}$$

where $\bar{C}_i^{\mu\nu}$ is the symmetric $p \times p$ correlation matrix given by

$$\bar{C}_i^{\mu\nu}(b) := N^{-1}\sum_{m\in V_i} \bar{x}_m^{\mu}\bar{x}_m^{\nu} \tag{24}$$

with $\mu, \nu = 1, \ldots, p$ and where

$$\Gamma_i^{\mu}(b) := \left[\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i) - \left(\sum_{m\in V_i^c} \hat{w}_{im}\bar{x}_m^{\mu} - \hat{\theta}_i\right)(2\xi_i^{\mu} - 1)\right](2\xi_i^{\mu} - 1). \tag{25}$$

From (23) we get, by multiplying both sides by $(2\xi_i^{\lambda} - 1)\bar{C}_i^{\mu\lambda}$ and summing over $\lambda = 1, \ldots, p$,

$$E_i^{\lambda}(b) = \sum_{\mu=1}^{p} \Gamma_i^{\mu}(b)(\bar{C}_i^{-1}(b))^{\mu\lambda}(2\xi_i^{\lambda} - 1) \tag{26}$$

where $\bar{C}^{-1}$ is the inverse of the matrix $\bar{C}$. With (26) we have obtained an alternative expression for the $E_i^{\mu}(b)$ (see equation (21)) in terms of the same quantities, namely $\hat{w}_{ij}$, ξ_i^{μ}, $\hat{\theta}_i$ and b. Substitution of this alternative expression (26) into (20) leads to an alternative expression for the $\hat{w}_{ij}$ with $j \in V_i$:

$$\hat{w}_{ij} = N^{-1} \sum_{\mu,\nu=1}^{p} \left[\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i) - \left(\sum_{m\in V_i^c} \hat{w}_{im}\bar{x}_m^{\mu} - \hat{\theta}_i\right)(2\xi_i^{\mu} - 1)\right](2\xi_i^{\mu} - 1)(\bar{C}_i^{-1}(b))^{\mu\nu}\bar{x}_j^{\nu}. \tag{27}$$

In equation (A.17) we introduced the $\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)$ as functions of the weights $\hat{w}_{ij}$. Here, we have found, conversely, the weights in terms of the $\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)$. By inserting $\hat{w}_{ij}$ (27) into $\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)$, equation (A.17), and making use of definition (24) for $\bar{C}_i^{\mu\nu}(b)$ one arrives, indeed, at an identity. In view of (17), equation (27) also holds true with all hats dropped.

The $\bar{\gamma}$ occurring in (27) are given by

$$\bar{\gamma}_i^{\mu}(b, \boldsymbol{w}_i) = N \sum_{\nu=1}^{p} \frac{\bar{C}_i^{\mu\nu}(b)(2\xi_i^{\nu} - 1)(2\xi_i^{\mu} - 1)\exp\left(-(\bar{\gamma}_i^{\nu}(b, \boldsymbol{w}_i))^2/2\sigma\right)}{\bar{\gamma}_i^{\nu}(b, \boldsymbol{w}_i)\sum_{\lambda}\exp\left(-(\bar{\gamma}_i^{\lambda}(b, \boldsymbol{w}_i))^2/2\sigma\right)} + \left(\sum_{m\in V_i^c} w_{im}\bar{x}_m^{\mu} - \theta_i\right)(2\xi_i^{\mu} - 1) \tag{28}$$

as follows from (21), (23) and (25). Equations (27) with (28) are an implicit expression for the weights. Developing the $\bar{\gamma}$ according to (22), we might obtain an explicit expression for the weights (27), just as in section 3.1.

The weights w_{ij} have been constructed as a solution of equation (9), an equation which is strongly related to equation (8). Hence, one may expect that, on the average, the γ_i^{μ} are positive, i.e.

$$\bar{\gamma}_i^{\mu}(b, \boldsymbol{w}_i) > 0. \tag{29}$$

We come now to the shortcut referred to above. Instead of determining the coefficients of the expansion (22) for the $\hat{\bar{\gamma}}$, we truncate this expansion after the first term. Dropping the hats and writing

$$\bar{\gamma}_i^{\mu 0} = \kappa \tag{30}$$

for *all* constant first terms in the expansions (22), we obtain from (27)

$$w_{ij} = \begin{cases} N^{-1} \displaystyle\sum_{\mu,\nu=1}^{p} \left[\kappa - \left(\sum_{m\in V_i^c} w_{im}\bar{x}_m^{\mu} - \theta_i\right)(2\xi_i^{\mu} - 1)\right](2\xi_i^{\mu} - 1) & \\ \quad\times(\bar{C}_i^{-1}(b))^{\mu\nu}\bar{x}_j^{\nu} & (j \in V_i) \\ w_{ij} \quad \text{(prescribed)} & (j \in V_i^c). \end{cases} \tag{31}$$

Note that with the choice $w_{ij}(t_0) = 0$ for $j \in V_i$ and $w_{ij}(t_0) = w_{ij}$ (prescribed) for $j \in V_i^c$ in our main result, equations (1) and (2), the latter equations reduce to equations (31). We thus have almost found the main result. The final form (1) and (2) is derived in section 5, after a numerical analysis of the particular case (31).

In view of (29), we will choose for κ, in equation (30), a certain positive number. This approach, in which we replace the constants $\bar{\gamma}_i^{\mu 0}$ by a number to be found by (numerical) trial and error is, *a priori*, rather crude. The usefulness of this way of handling will be the subject of the next section.

4. Numerical results: probing the basins

In this section we study the question regarding the size of the basins of attraction induced by the collection of patterns $\Omega^\mu(b)$. Stated differently, we will determine whether the solution (31) for the weights gives suitable basins of attraction. In particular, we will search for the optimal values κ and b to be taken in (31). This will be done by carrying out a numerical analysis.

Let us denote, more extensively, the γ_i^μ of equation (6) by $\gamma_i(\boldsymbol{x}, \boldsymbol{w}_i(b), \xi_i^\mu)$. Equation (5), with weights $w_{ij}(b)$ given by equations (31), is satisfied if, for a certain pattern $\boldsymbol{x}$, the $\gamma_i(\boldsymbol{x}, \boldsymbol{w}_i(b), \xi_i^\mu)$ are positive for all i. Therefore, we proceed as follows.

We construct probes consisting of patterns $\boldsymbol{x}$ centred around the typical patterns $\boldsymbol{\xi}^\mu$, and test whether these $\boldsymbol{x}$ are recognized by the neural net, i.e. we determine the sign of the γ_i for the patterns $\boldsymbol{x}$ of the probe. As a probing set we take patterns which are distributed around the typical patterns $\boldsymbol{\xi}^\mu$ in the same way as before, namely as given by formulae (10) and (11), but now with the basin parameter b replaced by a parameter $\bar{b}$. The latter parameter is dubbed the 'probing parameter'. In general, the probing parameter $\bar{b}$ used in the test will be unequal to the basin parameter b used to calculate the weights $w_{ij}(b)$. If the probing parameter $\bar{b}$ vanishes, a probing collection $\Omega^\mu(\bar{b} = 0)$ consists of precisely one pattern, namely $\boldsymbol{\xi}^\mu$.

In our numerical study, we first picked a certain value for the probing parameter $\bar{b}$, thereafter took an $\boldsymbol{x}$ belonging to the probing set $\Omega^\mu(\bar{b})$ defined by this $\bar{b}$, and thereupon calculated the $\gamma_i(\boldsymbol{x}, \boldsymbol{w}_i(b), \xi_i^\mu)$, equation (6). We repeated this procedure (for fixed $\bar{b}$) many times, and then calculated the fraction of $\boldsymbol{x}$ of the probing set for which all $\gamma_i(\boldsymbol{x}, \boldsymbol{w}_i(b), \xi_i^\mu)$ were positive.

In figure 1, we have depicted the relative number of $\boldsymbol{x}$ belonging to the basin (vertical axis) as a function of the basin parameter b (horizontal axis). The graphs a, b, c and d in figure 1 correspond to four values of the margin parameter κ: $\kappa = 1$, $\kappa = 2N^{-1}$, $\kappa = N^{-1}$ and $\kappa = \frac{1}{2}N^{-1}$. All patterns $\boldsymbol{\xi}^\mu$ are supposed to have the property that an arbitrary chosen ξ_i^μ has probability a to be equal to 1. This probability a is referred to as the mean activity. Note that for random patterns the mean activity is given by $a = 0.5$. Experimentally, however, the mean activity is found to be smaller [24]. In all graphs we have chosen vanishing prescribed weights, $w_{ij} = 0$, $j \in V_i^c$, and $\theta_i = N^{-1}$ for all $i = 1, \ldots, N$. That is, we considered diluted networks. More specifically, we took, randomly, 20% of the weights to belong to the set V_i^c, which corresponds to a dilution $d = 0.2$.

Each of the curves in figure 1 corresponds to a different value of the probing parameter $\bar{b}$. Going from top to bottom in the four graphs of figure 1, we cross curves with a larger and larger probing parameter $\bar{b}$. For the smallest possible value of the probing parameter $\bar{b}$, namely $\bar{b} = 0$, the probing set reduces to a typical pattern $\boldsymbol{\xi}^\mu$. It follows from figure 1 (see the upper curves, ◊——◊) that the fraction of $\boldsymbol{x}$ belonging to a basin equals 1 for a large range of the basin parameter b. As is to be expected, a typical pattern $\boldsymbol{\xi}^\mu$ indeed is a fixed point for all values of b (up to some upper limit which is larger than 0.3).

For values of the probing parameter $\bar{b}$ close to zero, $\bar{b} = 0.02$ say, the fraction of $\boldsymbol{x}$ belonging to a basin equals one for a large range of the basin parameter b (see the second curves from above, indicated by +· · ·+). As long as the probing set is smaller than the set

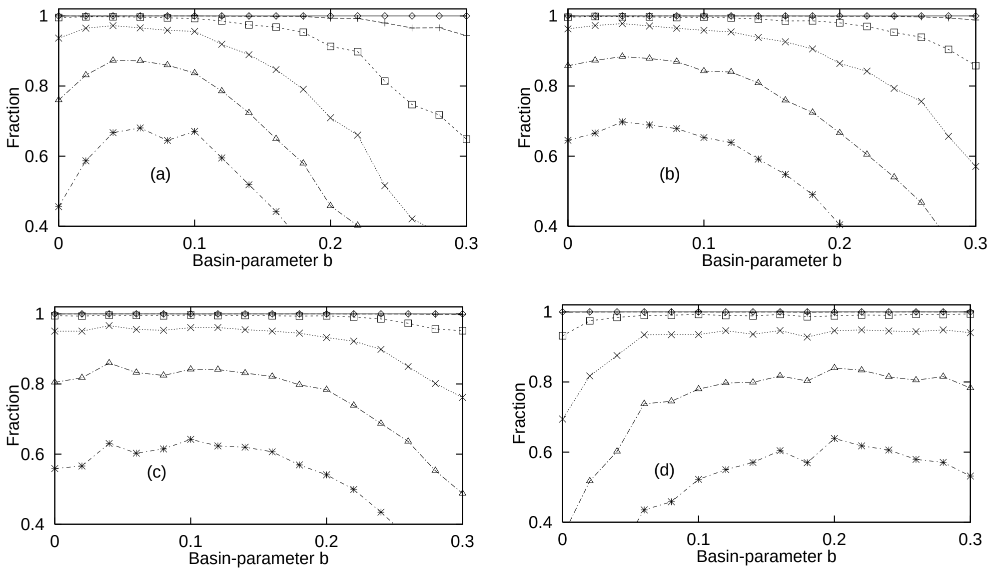

Figure 1. Probing of the basins for various values of the margin parameter. In the four graphs (a)–(d), the fraction of x with all γ_i positive is depicted, vertically, for four values of the parameter κ occurring in the final expression for the weights $\kappa = 1$, $\kappa = 2N^{-1}$, $\kappa = N^{-1}$ and $\kappa = \frac{1}{2}N^{-1}$ as a function of the basin parameter b. The six curves in each of the graphs correspond to different values of the probing parameter $\bar{b}$ that characterize the sets $\Omega^{\mu}(\bar{b})$. From top to bottom, in each graph, we have plotted the fraction of $\boldsymbol{x}$ with all γ_i positive for values of $\bar{b}$ given by the six numbers 0, 0.02, 0.04, 0.06, 0.08 and 0.1, respectively. The number of neurons is $N = 256$, the number of patterns $\boldsymbol{\xi}$ equals $p = 32$. The mean activity is $a = 0.2$. The dilution of the network is $d = 0.2$. In each of the four graphs, (a)–(d), that is, for four different values of the margin parameter κ, there is an interval of values of b for which the fraction of γ equals one, for a range of values of the probing parameter $\bar{b}$. Hence, for probes with $\bar{b}$ in the latter range, the net has values for the weights $w_{ij}(b)$ which are such that the net performs optimally.

of patterns which belong to the basin of attraction, the fraction remains one. If this fraction is less than one, the probing set is larger than the set of patterns which form the basins of attraction. Hence, the probing parameter $\bar{b}$ can be viewed as a measure for the size of the basin of attraction.

To illustrate these latter statements we take as an example figure 1(d). The lines $\bar{b} = 0$ and $\bar{b} = 0.02$ coincide: they are the horizontal line with fraction one. For $\bar{b} = 0.04$, corresponding to a fraction given by the curve with $\Box \cdots \Box$, the fraction rises to one as a function of b. This implies that the size of the basins grows as a function of b. For larger values of $\bar{b}$, given by the curves with crosses, triangles and asterisks, the fraction also rises as a function of b, up to some value of b, but never equals one. So in these cases, the number of elements of the probing sets always clearly is larger than the number of elements belonging to the basins.

Now we come to the effect of κ on the performance of the network. Comparing figures 1(a) and (d), and looking where the fraction equals one, we discover that for large κ, b should be small, and vice versa.

In figure 2, we study for a large value $\kappa = 1$ and a small value $\kappa = \frac{1}{2}N^{-1}$ of the margin parameter what happens when the number of patterns varies from 16 via 32 to 64. As before we have taken vanishing prescribed weights, $w_{ij} = 0$, $j \in V_i^c$, $\theta_i = N^{-1}$ for all $i = 1, \ldots, N$, and dilution $d = 0.2$. We find for $\kappa = 1$ as well as $\kappa = \frac{1}{2}N^{-1}$ that when the number of patterns increases, the size of the basins decreases. But, since the curves have a hump, a basin

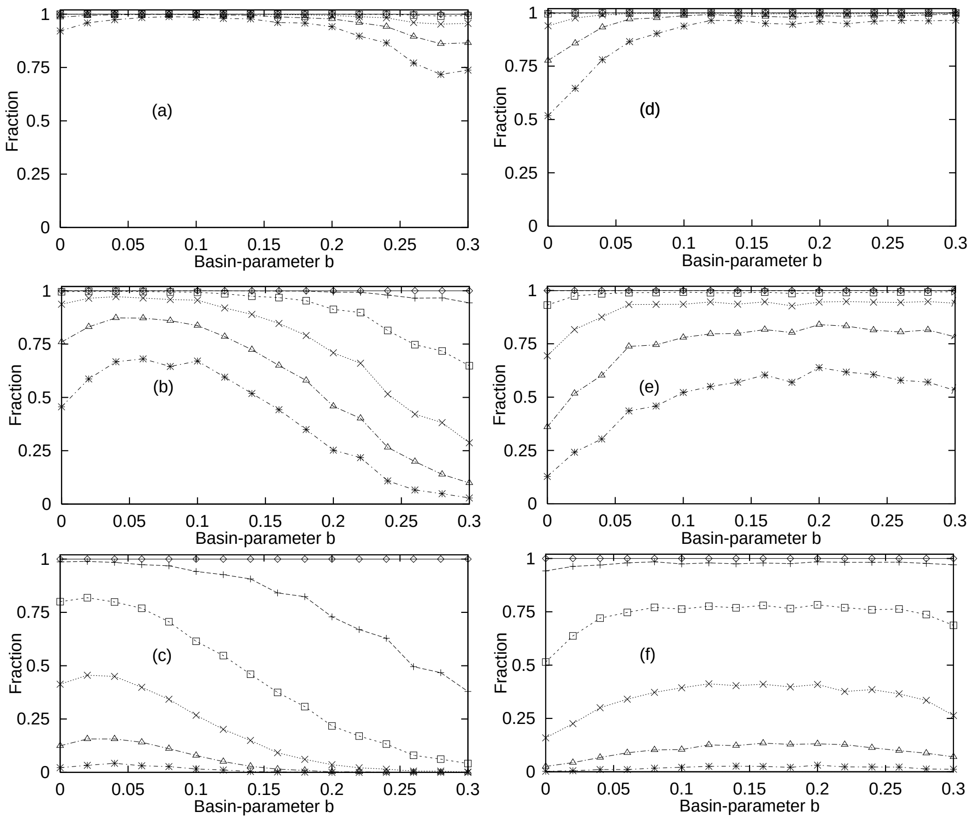

Figure 2. Probing of the basins for various numbers of patterns. The fraction of $\boldsymbol{x}$ with all γ_i positive is depicted, vertically, for three different values p of the number of stored patterns, $p = 16$ (top), $p = 32$ and $p = 64$ (bottom), as a function of the basin parameter b. In the left column the margin parameter is chosen large compared with the threshold, $\kappa = 1$, whereas in the right column κ is taken as of the order of the threshold, $\kappa = \frac{1}{2}N^{-1}$. The six curves in each of the graphs correspond to different values of the probing parameter $\bar{b}$. From top to bottom in each graph we have plotted the fraction of $\boldsymbol{x}$ with all γ_i positive for values of $\bar{b}$ given by 0, 0.02, 0.04, 0.06, 0.08 and 0.1, respectively. The number of neurons is $N = 256$, the mean activity is $a = 0.2$. The dilution of the network is $d = 0.2$. It is seen that for $b \neq 0$, the fraction rises, up to some value of b. Hence, for large κ (left column) and small κ (right column), the net performs better for $b \neq 0$, for different values of the number of patterns p.

parameter b with value unequal to zero yields a network that recognizes a larger part of the probing sets $\Omega^\mu(\bar{b})$.

A final observation relating to figures 1 and 2 is that, in general, a network with weights $w_{ij}(b \neq 0)$ possesses larger basins of attraction than a network with weights $w_{ij}(b = 0)$.

5. Relation to earlier work

The above mathematical study has been performed for adaptable weights, w_{ij}, $j \in V_i$, to be determined by the equations (9), and prescribed weights, w_{ij}, $j \in V_i^c$. Let us turn to the situation of a neural network that adapts its weights, in the course of time, according to some learning rule. In such a network, all weights start, at $t = t_0$ say, with some initial value $w_{ij}(t_0)$. The weights w_{ij}, with $j \in V_i^c$, keep their weights throughout the learning process, while the weights w_{ij}, with $j \in V_i$, change in the course of time. Now we ask the question, whether we can find $\tilde{w}_{ij}$ which are such that $\tilde{w}_{ij}(t)$ has prescribed values $w_{ij}(t_0)$, for all i and j, at $t = t_0$, whereas $\bar{\gamma}_i^\mu(b, \tilde{\boldsymbol{w}}_i(t))$ has a large probability of being positive. One way to obtain these $\tilde{w}_{ij}$ is via the w_{ij} that are given by the unhatted counterpart of equation (27). In fact, they are given by

$$\tilde{w}_{ij}(t) = \begin{cases} w_{ij}(t_0) + v_{ij}(t) & (j \in V_i) \\ w_{ij}(t_0) & (j \in V_i^c) \end{cases} \tag{32}$$

where

$$v_{ij}(t) = w_{ij}(t) - N^{-1} \sum_{\mu,\nu=1}^{p} \sum_{m \in V_i} w_{im}(t_0) \bar{x}_m^\mu \bar{C}_i^{-1}(b))^{\mu\nu} \bar{x}_j^\nu \tag{33}$$

in which we have denoted the (unhatted counterparts of) w_{ij} of equation (27) as $w_{ij}(t)$. An alternative way to write equation (33) is given by

$$v_{ij}(t) = N^{-1} \sum_{\mu,\nu=1}^{p} [\bar{\gamma}_i^\mu(b, \boldsymbol{w}_i(t)) - \bar{\gamma}_i^\mu(b, \boldsymbol{w}_i(t_0))](2\xi_i^\mu - 1)(\bar{C}_i^{-1}(b))^{\mu\nu} \bar{x}_j^\nu. \tag{34}$$

The weights $\tilde{w}_{ij}$, equation (32), have been constructed in such a way that

$$\bar{\gamma}_i^\mu(b, \tilde{\boldsymbol{w}}_i(t)) = \bar{\gamma}_i^\mu(b, \boldsymbol{w}_i(t)). \tag{35}$$

The latter equation can be verified easily. In fact, inserting (32) with (34) into (15) gives

$$\begin{aligned} \bar{\gamma}_i^\mu(b, \tilde{\boldsymbol{w}}_i(t)) = \bar{\gamma}_i^\mu(b, \boldsymbol{w}_i(t_0)) + \sum_{\nu,\lambda=1}^{p} [\bar{\gamma}_i^\nu(b, \boldsymbol{w}_i(t)) - \bar{\gamma}_i^\nu(b, \boldsymbol{w}_i(t_0))](2\xi_i^\nu - 1) \\ \times (2\xi_i^\mu - 1)(\bar{C}_i^{-1}(b))^{\lambda\nu} \bar{C}_i^{\lambda\mu}(b) \end{aligned} \tag{36}$$

where we used definitions (15) and (24). Since $\bar{C}_i^{\lambda\mu}(b)$ is symmetric, the product of the matrices $\bar{C}$ gives a Kronecker delta, which in turn yields (35). The property (35) guarantees that when the $\bar{\gamma}_i^\mu(b, \boldsymbol{w}_i(t))$ are positive, the $\bar{\gamma}_i^\mu(b, \tilde{\boldsymbol{w}}_i(t))$ are also positive.

Using the same shortcut as above, equation (30), we obtain

$$v_{ij}(t) = N^{-1} \sum_{\mu,\nu=1}^{p} [\kappa - \bar{\gamma}_i^\mu(b, \boldsymbol{w}_i(t_0))](2\xi_i^\mu - 1)(\bar{C}_i^{-1}(b))^{\mu\nu} \bar{x}_j^\nu \tag{37}$$

with $\bar{x}_j^\nu$ given by (14). Equations (32) and (37) are equivalent to the main result (1) and (2) mentioned in the introduction. Putting in this expression the basin parameter equal to zero ($b = 0$), we recover the expression obtained after a learning process in a preceding paper [21]. This suggest that (32) with (37) is the generalization of the weights in a process of learning with noisy patterns. Hence, we may state that a network performs optimally when trained with noise ($b \neq 0$), or, stated differently (and less precisely), a neural network performs best in an environment identical to the training environment. This is what Wong and Sherrington refer to as the 'principle of adaptation' [4]. In our next paper, we will extensively return to this question, in a biological context [22]. The final result will turn out to be that expression (32) with (37) is, apart from a detail, indeed the generalization of learning with noisy patterns.

6. Conclusion

Although we studied a neural network, we did not consider learning and learning rules. We simply asked the question, what values must one take for the weights of a neural network in order that it performs optimally, i.e. that it can retrieve the largest sets of perturbed patterns.

We were able to reformulate this problem in a mathematically exact way, and to obtain a solution that, by its construction, had a certain plausibility of being a suitable one. Finally, we performed a numerical test, which confirmed the usefulness of our approach. The weights $w_{ij}(b)$ obtained in this paper on the basis of perturbed data ($b \neq 0$) yield a network with larger basins than would have been obtained in the case of non-perturbed data ($b = 0$). In a subsequent paper we will propose a biological learning rule which is such that, apart from a minor detail, the synapses strive at the values for their weights as given by the main result of this paper, equations (1) and (2). In other words, nature might realize almost totally what mathematics suggests.

Appendix A. Derivation of implicit equations for the weights

In this appendix we will evaluate the left-hand side of equation (9). Then, combining this with the result of section 3.1 for the right-hand side will lead to implicit equations for $\hat{w}_{ij}$.

Inserting (19) into the left-hand side of (9), multiplying by a delta function containing a variable z and integrating over z, we get the equivalent expression

$$\sum_{\mu}\sum_{x_1=0,1}\cdots\sum_{x_N=0,1} p^{\mu}(\boldsymbol{x})\int \mathrm{d}z x_j \Theta_{\mathrm{H}}(\hat{w}_{ij}x_j - \hat{\theta}_i + z)\delta\left[z - \sum_{l\neq j}\hat{w}_{il}x_l\right]$$
$$= \sum_{\mu}\int \mathrm{d}z \sum_{x_j} p_j^{\mu}(x_j)x_j\Theta_{\mathrm{H}}(\hat{w}_{ij}x_j - \hat{\theta}_i + z)P_{ij}^{\mu}(z) \tag{A.1}$$

where we used (10) and where we abbreviated

$$P_{ij}^{\mu}(z) = \sum_{x_1}\cdots\sum_{x_{j-1}}\sum_{x_{j+1}}\cdots\sum_{x_N}\prod_{m\neq j} p_m^{\mu}(x_m)\delta\left[z - \sum_{l\neq j}\hat{w}_{il}x_l\right] \qquad (j \in V_i). \tag{A.2}$$

The summation over x_j in (A.1) yields

$$\sum_{x_j} p_j^{\mu}(x_j)x_j\Theta_{\mathrm{H}}(\hat{w}_{ij}x_j - \hat{\theta}_i + z) = \bar{x}_j^{\mu}\Theta_{\mathrm{H}}(\hat{w}_{ij} - \hat{\theta}_i + z) \tag{A.3}$$

as follows by inserting (11). The factor $P_{ij}^{\mu}(z)$ can be rewritten in the following way.

Using a well known representation of the delta function we first obtain

$$P_{ij}^{\mu}(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}k\mathrm{e}^{\mathrm{i}kz}\prod_{m\neq j}\sum_{x_m} p_m^{\mu}(x_m)\mathrm{e}^{-\mathrm{i}k\hat{w}_{im}x_m}. \tag{A.4}$$

One has

$$\sum_{x_m} p_m^{\mu}(x_m)\mathrm{e}^{-\mathrm{i}k\hat{w}_{im}x_m} = (1-b)\mathrm{e}^{-\mathrm{i}k\hat{w}_{im}\xi_m^{\mu}} + b\mathrm{e}^{-\mathrm{i}k\hat{w}_{im}(1-\xi_m^{\mu})} \tag{A.5}$$

where we used (11). Inserting (A.5) into (A.4) we may write

$$P_{ij}^{\mu}(z) = \frac{1}{2\pi}\int \mathrm{d}k \exp\left\{\mathrm{i}kz + \sum_{m\neq j}\ln\left[(1-b)\mathrm{e}^{-\mathrm{i}k\hat{w}_{im}\xi_m^{\mu}} + b\mathrm{e}^{-\mathrm{i}k\hat{w}_{im}(1-\xi_m^{\mu})}\right]\right\} \tag{A.6}$$

where we used $y = \exp\{\ln y\}$. We can now expand the two exponentials occurring in the argument of the logarithm. This leads to a term of the form $\ln(1 + y)$. Thereupon, we can expand this term as $y - \frac{1}{2}y^2 + \cdots$, since y is of the order of $\hat{w}_{ij}$, and $\hat{w}_{ij}$ is of the order $N^{-1/2}$, as noted above (see equations (17) and following text). Thus we obtain

$$\ln\left[(1-b)\mathrm{e}^{-\mathrm{i}k\hat{w}_{im}\xi_m^{\mu}} + b\mathrm{e}^{-\mathrm{i}k\hat{w}_{im}(1-\xi_m^{\mu})}\right] = -\mathrm{i}k\hat{w}_{im}\bar{x}_m^{\mu} - \tfrac{1}{2}b(1-b)k^2\hat{w}_{im}^2 + \cdots. \tag{A.7}$$

Inserting (A.7) into (A.6) we may write

$$P_{ij}^{\mu}(z) = \frac{1}{2\pi} \exp\{-(z-z_0)^2/2\sigma\} \int_{-\infty}^{\infty} \mathrm{d}k \exp\left\{-\frac{\sigma}{2}(k - \mathrm{i}(z-z_0)/\sigma)^2\right\} + \cdots \tag{A.8}$$

where we abbreviated

$$\sigma := b(1-b) \sum_{m \neq j} \hat{w}_{im}^2 \qquad z_0 := \sum_{m \neq j} \hat{w}_{im} \bar{x}_m^{\mu}. \tag{A.9}$$

Using the fact that $\hat{w}_{ij}$ is of the order $1/\sqrt{N}$ we may write

$$\sigma = b(1-b) \tag{A.10}$$

a relation we will use later. After evaluating the integral (A.8), we obtain

$$P_{ij}^{\mu}(z) = (2\pi\sigma)^{-\frac{1}{2}} \exp\{-(z-z_0)^2/2\sigma\} + \cdots \qquad (i = 1, \ldots, N;\ j \in V_i) \tag{A.11}$$

with $\mu = 1, \ldots, p$. Substituting (A.3) and (A.11) into the right-hand side of (A.1) we obtain for the left-hand side of (9)

$$(2\pi\sigma)^{-\frac{1}{2}} \sum_{\mu} \int \mathrm{d}z \bar{x}_j^{\mu} \Theta_{\mathrm{H}}(\hat{w}_{ij} - \hat{\theta}_i + z) \exp\{-(z-z_0)^2/2\sigma\}. \tag{A.12}$$

The integral occurring in (A.12) can be rewritten

$$I_{ij}^{\mu} := (2\pi\sigma)^{-\frac{1}{2}} \int \mathrm{d}z \Theta_{\mathrm{H}}(\hat{w}_{ij} - \hat{\theta}_i + z) \exp\{-(z-z_0)^2/2\sigma\}. \tag{A.13}$$

Changing the integration variable z according to $y = (z-z_0)/\sqrt{2\sigma}$, we find

$$\begin{aligned} I_{ij}^{\mu} = \pi^{-\frac{1}{2}} \int_{-\infty}^{\infty} \mathrm{d}y \Theta_{\mathrm{H}}\left(\hat{w}_{ij} - \hat{\theta}_i + z_0 + \sqrt{2\sigma} y\right) \mathrm{e}^{-y^2} \\ = \pi^{-\frac{1}{2}} \int_0^{\infty} \mathrm{d}y \mathrm{e}^{-y^2} + (4\pi)^{-\frac{1}{2}} \int_0^{\infty} \mathrm{d}y \Big[\mathrm{sgn}\left(\hat{w}_{ij} - \hat{\theta}_i + z_0 - \sqrt{2\sigma} y\right) \\ + \mathrm{sgn}\left(\hat{w}_{ij} - \hat{\theta}_i + z_0 + \sqrt{2\sigma} y\right) \Big] \mathrm{e}^{-y^2}. \end{aligned} \tag{A.14}$$

The integral over the first term is a Gaussian integral; the second term can be expressed in an error function. We obtain

$$I_{ij}^{\mu} = \tfrac{1}{2} + \tfrac{1}{2} \mathrm{erf}\left([\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)(2\xi_i^{\mu} - 1) + \epsilon_{ij}^{\mu}]/\sqrt{2\sigma}\right) \qquad (i = 1, \ldots, N;\ j \in V_i) \tag{A.15}$$

where $\mu = 1, \ldots, p$ and where the error function is defined according to

$$\mathrm{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{d}y \mathrm{e}^{-y^2}. \tag{A.16}$$

In analogy to (15) we defined

$$\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i) = \left(\sum_{l=1}^{N} \hat{w}_{il} \bar{x}_l^{\mu} - \hat{\theta}_i \right)(2\xi_i^{\mu} - 1). \tag{A.17}$$

Furthermore, we abbreviated

$$\epsilon_{ij}^{\mu} = -\hat{w}_{ij} \bar{x}_j^{\mu} + \hat{w}_{ij}. \tag{A.18}$$

Note that, apart from a $\boldsymbol{\xi}^{\mu}$-dependent factor, the quantity ϵ_{ij}^{μ} equals the weight $\hat{w}_{ij}$. In view of (17), $\epsilon_{ij}^{\mu}/\sqrt{2\sigma}$ is small. The error function in (A.15) can be split into two contributions. For small ϵ we have

$$\int_{\gamma}^{\gamma+\epsilon} \mathrm{d}y \mathrm{e}^{-y^2} = \epsilon \mathrm{e}^{-\gamma^2} + \cdots \tag{A.19}$$

which allows us to write for (A.15)

$$I_{ij}^{\mu} = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)(2\xi_i^{\mu} - 1)/\sqrt{2\sigma}\right) + \frac{\epsilon_{ij}^{\mu}}{\sqrt{2\pi\sigma}} \exp\left(-(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i))^2/2\sigma\right) + \cdots . \tag{A.20}$$

Using (A.12) and (A.20) with (A.17), the final expression for the left-hand side of (9) can be obtained:

$$\frac{1}{2}\sum_{\mu} \bar{x}_j^{\mu}\left[1 + \mathrm{erf}\left(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)(2\xi_i^{\mu} - 1)/\sqrt{2\sigma}\right)\right] + \frac{b(1-b)\hat{w}_{ij}}{\sqrt{2\pi\sigma}} \sum_{\mu} \exp\left(-(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i))^2/2\sigma\right). \tag{A.21}$$

Combining the right- and left-hand sides of equation (9), as given by (16) and (A.21), respectively, we get an equation from which the weights $\hat{w}_{ij}$ follow immediately:

$$\hat{w}_{ij} = \frac{\sqrt{2\pi\sigma}\sum_{\mu} \bar{x}_j^{\mu}\left[(2\xi_i^{\mu} - 1) - \mathrm{erf}\left(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)(2\xi_i^{\mu} - 1)/\sqrt{2\sigma}\right)\right]}{2b(1-b)\sum_{\mu} \exp\left(-(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i))^2/2\sigma\right)}. \tag{A.22}$$

With the properties

$$\mathrm{erf}\left(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)(2\xi_i^{\mu} - 1)/\sqrt{2\sigma}\right) = (2\xi_i^{\mu} - 1)\mathrm{erf}\left(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)/\sqrt{2\sigma}\right) \tag{A.23}$$

and

$$\mathrm{erf}(y) = 1 - \frac{1}{y\sqrt{\pi}}\mathrm{e}^{-y^2} + \cdots \tag{A.24}$$

we can rewrite (A.22),

$$\hat{w}_{ij} = \frac{\sqrt{2\pi\sigma}}{2b(1-b)} \sum_{\mu} \bar{x}_j^{\mu}(2\xi_i^{\mu} - 1)\left[\sqrt{\pi/2\sigma}\,\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i)\right]^{-1} \times \exp\left(-(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i))^2/2\sigma\right) \Big/ \sum_{\mu} \exp\left(-(\hat{\bar{\gamma}}_i^{\mu}(b, \boldsymbol{w}_i))^2/2\sigma\right) \tag{A.25}$$

or, equivalently, the final results (20) with (21) of the main text.

References

[1] Bastolla U and Parisi G 1997 *J. Phys. A: Math. Gen.* **30** 5613
[2] Wong K Y M and Ho C 1994 *J. Phys. A: Math. Gen.* **27** 5167
[3] Wong K Y M 1993 *Physica* A **200** 619
[4] Wong K Y M and Sherrington D 1992 *Physica* A **185** 453
[5] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** 4659
[6] Amit D J, Evans M R, Horner H and Wong K Y M 1990 *J. Phys. A: Math. Gen.* **23** 3361
[7] Wong K Y M and Sherrington D 1993 *Phys. Rev.* E **47** 4465
[8] Wong K Y M and Sherrington D 1990 *J. Phys. A: Math. Gen.* **23** L175
[9] Gardner E J, Stroud N and Wallace D J 1989 *J. Phys. A: Math. Gen.* **22** 2019
[10] Kitano K and Aoyagi T 1998 *J. Phys. A: Math. Gen.* **31** L613
[11] Rodrigues Neto C and Fontanari J F 1996 *J. Phys. A: Math. Gen.* **29** 3041
[12] Erichsen R Jr and Theumann W K 1995 *Physica* A **220** 390
[13] Yau H W and Wallace D J 1991 *J. Phys. A: Math. Gen.* **24** 5639
Yau H W and Wallace D J 1992 *Physica* A **185** 471
[14] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969
[15] Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949
[16] Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745

[17] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[18] Kepler T B and Abbott L F 1988 *J. Phys., Paris* **49** 1657
[19] Forest B M 1988 *J. Phys. A: Math. Gen.* **21** 245
[20] Rodrigues Neto C and Fontanari J F 1997 *J. Phys. A: Math. Gen.* **30** 7945
[21] Heerema M and van Leeuwen W A 1999 *J. Phys. A: Math. Gen.* **32** 263
[22] Heerema M and van Leeuwen W A 2000 *J. Phys. A: Math. Gen.* **33** 1781 (following article)
[23] Wiegerinck W and Coolen A 1993 *J. Phys. A: Math. Gen.* **26** 2535
[24] Abelles M 1982 *Studies of Brain Function* (New York: Springer)