Derivation of Hebb's rule

1999 J. Phys. A: Math. Gen. 32 263

(http://iopscience.iop.org/0305-4470/32/2/004)

J. Phys. A: Math. Gen. **32** (1999) 263–286. Printed in the UK PII: S0305-4470(99)97686-2

Derivation of Hebb's rule

M Heerema† and W A van Leeuwen‡

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

Received 5 November 1997, in final form 16 September 1998

Abstract. On the basis of the general form for the energy needed to adapt the connection strengths w_{ij} of a network in which learning takes place, a local learning rule is found for the changes Δw_{ij}. This biologically realizable learning rule turns out to comply with Hebb's neuro-physiological postulate, but is not of the form of any of the learning rules proposed in the literature.

The learning rule possesses the property that the energy needed in each learning step is minimal, and is, as such, evolutionary attractive. Moreover, the pre- and post-synaptic neurons are found to influence the synaptic changes differently, resulting in an asymmetric connection matrix w_{ij}, a fact which is in agreement with biological observation.

It is shown that if a finite set of the same patterns is presented over and over again to the network, the weights of the synapses converge to finite values.

Furthermore, it is proved that the final values found in this biologically realizable limit are the same as those found via a mathematical approach to the problem of finding the weights of a partially connected neural network that can store a collection of patterns. The mathematical solution is obtained via a modified version of the so-called method of the pseudo-inverse, and has the inverse of a reduced correlation matrix, rather than the usual correlation matrix, as its basic ingredient. Thus, a biological network might realize the final results of the mathematician by the energetically economic rule for the adaption of the synapses found in this article.

1. Introduction

In this paper we consider some theoretical aspects of the changes of the connections as they could take place between the nerve cells, or neurons, of the brain. In a learning process, these connections change continuously, and are adapted in such a way that a particular task, e.g. the storage of patterns, is achieved. The answer to the question in which way the connections between neurons actually change in response to external stimuli, can only be given by experiment, not via any theoretical discussion. Although there is a lot of experimental activity related to the study of the functioning of neurons, there is not yet a unique answer to this question: see, e.g. the 1998 review articles of Buonomano and Merzenich [1], Marder [2], or the 1990 review article of Brown *et al* [3].

In the 1940s, the Canadian psychologist Hebb conjectured (in his well known book *The organization of behaviour—A neuro-physiological theory* [4]) that the changes of the connections between the neurons take place according to a 'neuro-physiological postulate' that nowadays is referred to as Hebb's rule: 'When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells so that A's efficiency, as one of the cells firing B, is increased'.

† E-mail address: heerema@phys.uva.nl
‡ E-mail address: leeuwen@phys.uva.nl

0305-4470/99/020263+24$19.50

Thus Hebb's rule is a quantitative statement on the enhancement of synaptic efficiency of signal transmission, but does not state qualitatively, by some mathematical formula, to what extent.

Nowadays, there is plenty of evidence that synapses do indeed change in a learning process and since the appearance of Hebb's article many quantitative proposals, all complying with Hebb's postulate, have been put forward. This paper is also concerned with such a quantitative expression for the synaptic changes. However, rather than postulating a learning rule, we derive it from some underlying principle. As a final result, we find a learning rule for the adaptation of the strengths, or weights, w_{ij}, of a synapse connecting a post-synaptic neuron i and a pre-synaptic neuron j. Its explicit form reads:

$$\Delta w_{ij}(t_n) = \eta_i[\kappa - \{h_i(t_n) - \theta_i\}(2\xi_i - 1)](2\xi_i - 1)\xi_j. \tag{1}$$

This—asymmetric—learning rule gives Δw_{ij}, the positive or negative increment of the weight w_{ij}, as a function of the activities ξ_i and ξ_j of neurons i and j of the synapse that connects these neurons. In our convention, the activity ξ of a neuron equals 1 if it generates an action potential, and 0 if it is quiescent. The function h is the potential difference between the interior and the exterior of a neuron, at its axon hillock. The formula gives the change at time t_n. The index n denotes the time at the nth learning step in the process of learning ($n = 1, 2, \ldots$). The threshold potential, θ_i, is a constant, typical for the neuron i in question. It equals, by definition, the potential that must be surmounted, at the axon hillock of neuron i, in order that it will fire. The quantities η_i and κ are also constants. Their precise identification, as variables related to individual and collective neuron properties, is outside the scope of this paper. The learning rule (1), which constitutes our main result as far as biology is concerned, has a form that is compatible with Hebb's postulate.

It is a well known fact that, for a given neural net with strengths w_{ij} of the weights, there are *infinitely* many ways to choose changes Δw_{ij} of the weights such that the network will perform storage and retrieval of a new pattern. The derivation of our learning rule is based on the assumption that, at each instant of the learning process, the energy needed to change the neural network in order to store a new pattern, is minimal. The requirement that, at each step n of the learning process, the energy needed is as low as possible, turns out to be sufficient to *uniquely* determine the way in which the weight of each synapse connecting two arbitrary neurons i and j should be changed, and thus fixes a learning rule for the adaptation of the weights of all the connections. We will call this learning rule the 'non-local energy saving learning rule', since it turns out to depend on the state of activity of *all* neurons j from which neuron i receives its input. It is given by equation (42) below.

It is impossible, however, for a synapse connecting two neurons i and j, to realize the non-local energy saving learning rule (42) exactly, as follows by a careful inspection of formula (42). In fact, in order to adapt itself according to this learning rule, a synapse between i and j would have to 'know' the individual states of activity ξ_k of *all* pre-synaptic neurons k from which neuron i gets its input, whereas a synapse only 'feels' the states of the two neurons i and j which it connects. The best a synapse can do in order to compete with the performance of the non-local learning rule (42) is to adapt itself according to a learning rule that is a local approximation of the non-local learning rule. It is this local approximation, given the expression (1) above, which constitutes our main biological result. We will refer to it as the local energy saving learning rule, to distinguish it from its non-local counterpart. The point of locality of learning rules is discussed in more detail in section 6.

A numerical estimation of the performance of the local learning rule, equation (1), versus to the non-local one, equation (42), is made in section 7. Local learning turns out to be a very effective alternative to non-local learning, regarding both its power to store and retrieve patterns and its capacity to be economic in use of energy.

In order to arrive at the non-local energy saving learning rule, we think of a neuron as a living cell. A living cell, as a physical object, is a stationary non-equilibrium system. The basic assumption of this paper is that any type of change of the cellular cleft can only be effected by *adding* energy to the non-equilibrium system, independent of whether it leads to a strengthening or a weakening of the synaptic efficacy. This is a plausible, but not totally trivial postulate, which can only be falsified by a detailed biophysical or biochemical study of the process of change of the synapse. In our setup, the mere assumption that extra energy is needed for any change of the synapse, independent of whether it leads to an increase or a decrease of its efficiency, replaces Hebb's postulate on efficiency cited above.

Before starting the derivation of the energy saving learning rule itself, we discuss, in section 3, the 81 possibilities which, in principle, are compatible with Hebb's postulate. In particular, we consider these mathematical realizations with respect to their biological plausibility. We then find that, in fact, out of the 81 learning rules that are possible in principle, only two are also biologically plausible. These are the learning rules (20) and (21).

The actual derivation of the energy saving learning rule is performed in section 4. To our satisfaction, it general form turns out to imply the two forms (20) and (21) expected in the preceeding section on biological grounds only. Thus our 'principle of minimal change of energy', which might lead, *a priori*, to any of the 81 possibilities for a realization of a learning rule for the change of weight of a synapse, happens to yield precisely those rules which are biological plausible.

In section 5 we consider the situation that the changes of the connections do not take place in an energetically optimal way, but in such a way that patterns are not partially wiped out when new patterns are learned as is the case for learning based on the energy saving learning rule (1) or (42). We then ask ourselves the question: which learning rule would then be found for the changes Δw_{ij} of the synaptic weights? Again, its general form turns out to comply with one of the 81 possible realizations of the Hebb rule considered in section 3, but, in this case, it is a biologically improbable one. We therefore do not pursue this path any further.

The question might arise whether the non-local energy saving learning rule converges, in the limit that the number of learning steps tends to infinity. And, if so, to what values they then would converge. The answers to these questions are the subject of section 4.2.

There exists a well known way to obtain the final form of the connection strengths w_{ij} of an artificial neural network that can store and retrieve a set of patterns: it goes under the name 'pseudo-inverse solution' [5, 6]. By inversion of a certain matrix related to the patterns to be stored, the so-called correlation matrix, one can obtain, without any limiting procedure, final values for the weights w_{ij} of the connections of a neural network that yield the desired result of being capable of storing and retrieving a collection of patterns.

We will consider an assembly of N neurons, where N is a number relevant for a certain subunit of the brain, such as a cortical hyper-column, for which N is of the order of 10^4–10^5. Although such subunits are highly interconnected, they are partially connected in the mathematical sense, since each neuron is connected to only a finite fraction of the subunit considered. Moreover, biological neurons are not self-connected, i.e. $w_{ii} = 0$. These two biological facts force us to study, from the very beginning, diluted, or partially connected, networks. In the limit that the dilution tends to zero, we rediscover, if we relax the requirement that the self-connections all vanish, some of the well known results for fully connected networks, in particular those of Diederich and Opper [7], and of Linkevich [8].

A possible question one might now ask is: is there any relation between the final values obtained for the weights w_{ij} obtained in the limit of an infinite number of learning steps, $n \to \infty$, on the one hand and the values obtained via the pseudo-inverse method on the other hand? The answer to this question is as simple as it is amazing: the results are identical. The

proof of this point is the subject of appendix B, where the method of the pseudo-inverse is modified in such a way that it can be used for partially connected networks. Thus, as a final conclusion, we can state that: (i) the assumption of economy of energy in a learning step, (ii) the well known method based on the pseudo-inverse of the correlation matrix and (iii) the biological plausibility of a learning rule are three members of a trio that work in concert. We want to stress, once again, that the question of whether the evolutionary development of the brain actually has led to an adaptation process of the synapses that is energetically the most economical, is, as yet, experimentally, an open question. It is not excluded that the realization of the changes of the synapses might take place in a biologically less probable, or an energetically less favourable way. Our only certainty is that economy of energy and biological probability go hand in hand.

Usually, neural networks have been modelled in the so-called spin representation, which, in principle, can easily be translated to the so-called binary representation, which models the biological reality more directly. In particular, in the binary representation the thresholds for activation of a neuron can be taken constant, in accordance with the biological reality. In the spin representation, however, the actual biological reality in a learning process can only be modelled via the use of a time-dependent threshold, a fact which is often overlooked: one erroneously treats the neuron thresholds in the spin model as constants (see, e.g. [9, 10]). We therefore have chosen not to use the spin, but the binary representation.

In our study of the connections w_{ij} and the way in which they change in a learning process, we will neglect two constraints set by nature. Firstly, the fact that, for an actual neuron i, the magnitudes of the synaptic connections are within some interval characteristic for the synapse in question. Secondly, the fact that, according to Dale's law, the connections related to one and the same pre-synaptic neuron either are only excitatory or only inhibitory. Furthermore, we treat biological neurons as McCulloch and Pitts neurons, i.e. their response to input is according to the rule (2), (3) below. We thus also neglect the retardation which results from the finite speed of transmission of signals through axons and dendrites. A way retardation could be included in a model has been put forward in [11].

For an introduction to this paper, see textbooks such as [12–14].

2. Attractor neural network model

Dynamics. We consider a network of N interconnected neurons in the binary representation, i.e. each neuron can have a state $x_i = 1$ (the neuron produces *one* action-potential or *spike*) or $x_i = 0$ (the neuron is quiescent). The post-synaptic potential of neuron i at time t of this system of neurons is modelled by

$$h_i(t) = \sum_{j=1}^{N} w_{ij}(t)x_j(t) \qquad (i = 1, \ldots, N) \tag{2}$$

where the $x_j(t)$ are the input signals at time t and where the $w_{ij}(t)$ are the *weights*, also called *synaptic strengths* or *synaptic efficacies* at time t. A weight w_{ij} takes into account the overall effect of a synaptic connection between a post-synaptic neuron i and a pre-synaptic neuron j and may be positive (excitation), negative (inhibition) or zero (no synaptic connection). The weights w_{ij}, like the potentials h_i, are expressed in volts. The output of neuron i is supposed to be given by the dynamical equation

$$x_i(t + \Delta t) = \theta_{\mathrm{H}}\{h_i(t) - \theta_i\} \qquad (i = 1, \ldots, N) \tag{3}$$

where the *constant* θ_i is the activation threshold characteristic of neuron i and where Δt is some discrete time step. A typical value for θ_i is 10 mV [15]. The symbol θ_{H} stands for the

Heaviside step function, which equals one for positive arguments and vanishes otherwise.

In the so-called 'spin representation', active and non-active states of neuron i are characterized by $s_i = 1$ or $s_i = -1$, respectively. In this representation, the dynamical equation (3) can be rewritten as

$$s_i(t + \Delta t) = \mathrm{sgn}\left\{ \sum_{j=1}^{N} J_{ij}(t) s_j(t) - T_i(t) \right\} \qquad (i = 1, \ldots, N) \tag{4}$$

where the time-dependent 'coupling constants' J_{ij} are related to the biological weights w_{ij} through $J_{ij} = w_{ij}/2$ and where $s_j = 2x_j - 1$. The time-dependent 'thresholds' $T_i(t)$ are related to the constant biological thresholds θ_i according to

$$T_i(t) = \theta_i - \sum_{j=1}^{N} J_{ij}(t) \qquad (i = 1, \ldots, N). \tag{5}$$

In the literature the thresholds $T_i(t)$ are usually treated as a constant; most often the constant is taken to vanish [9, 10]. This seemingly innocent fact changes, of course, the dynamics (4) of the system in a non-trivial way. As a consequence, the results obtained for, e.g. the adaptation of the coupling constants differ from those obtained when the actual biological dynamics (3) is used (cf equations (44) and (45)). Hence, when modelling adaptation processes of biological neurons with constant thresholds, the use of the binary representation is obligatory.

Neural networks have two timescales, one related to the rate of change of the synaptic efficacies w_{ij} and one related to the spiking activity of a neuron. The latter time is of the order of milliseconds, the former is less well-defined, but can be estimated to lie somewhere between seconds and days: it is a time related to the rate of learning of a brain. Hence, the Δt occurring in equation (3) is of the order of milliseconds. When the process of adaptation of the weights has come to an end the w_{ij} remain constant.

Fixed points. We want to determine the synaptic efficacies of an attractor neural network, i.e. of a network which can recall a number, p say, of previously stored patterns. The realization of a recall corresponds to a fixed network state of the network dynamics (3). Let us denote the patterns of activity, or patterns, by $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)$, where $\mu = 1, \ldots, p$. Thus $\xi_i^\mu = 1$ or $\xi_i^\mu = 0$ with $i = 1, \ldots, N$ and $\mu = 1, \ldots, p$. The probability that a neuron i is in the state 1 or 0 is supposed to be given by a or $(1 - a)$ respectively. The quantity a is usually called the mean activity of the neural net. For random patterns the mean activity a is given by 0.5. In biological neural networks, however, the mean activity a is smaller [16].

Thus, a network which has stored, somehow, p patterns $\boldsymbol{\xi}^\mu$ satisfies the fixed point equations

$$x_i(t + \Delta t) = x_i(t) \qquad \text{for} \quad x_i(t) = \xi_i^\mu \qquad (i = 1, \ldots, N; \mu = 1, \ldots, p). \tag{6}$$

Hence, equations (3) and (6) yield the pN equations

$$\xi_i^\mu = \theta_{\mathrm{H}}\left\{ \sum_{j=1}^{N} w_{ij}(t) \xi_j^\mu - \theta_i \right\} \tag{7}$$

for N^2 unknown w_{ij}.

Let us now introduce so-called stability coefficients γ_i^μ [17]:

$$\gamma_i^\mu(t) := (h_i^\mu(t) - \theta_i)(2\xi_i^\mu - 1) \tag{8}$$

with h_i^μ the post-synaptic potential

$$h_i^\mu(t) = \sum_{j=1}^{N} w_{ij}(t) \xi_j^\mu. \tag{9}$$

Remark that γ_i^μ depends, via h_i^μ, on all weights w_{ij}, i.e.

$$\gamma_i^\mu(t) = \gamma_i^\mu(w_{11}(t), w_{12}(t), \ldots, w_{N-1,N}(t), w_{NN}(t)).$$

One easily checks, by distinguishing the cases $\xi_i^\mu = 1$ and $\xi_i^\mu = 0$, that an equivalent way to express the equalities (7) are the pN *in*equalities

$$\gamma_i^\mu(t) > 0. \tag{10}$$

The inequality sign in (10) reflects that fact that the set of equations (7) is under-determined, i.e. the equations (10) are necessary but not sufficient equations to determine uniquely a set of weights of a network which has stored some patterns.

An arbitrary pattern $\boldsymbol{X}(t)$ will only be recalled if it evolves in time to one of the fixed points $\boldsymbol{\xi}^\mu$. Therefore, it is not sufficient for a network to have fixed points: for each of the p fixed points that is related to a retrieval of a pattern $\boldsymbol{\xi}^\mu$, there must exist a whole neighbourhood of points around $\boldsymbol{\xi}^\mu$ which is such that all points of this neighbourhood will evolve to $\boldsymbol{\xi}^\mu$ under the dynamics (3). In technical terms, the fixed points $\boldsymbol{\xi}^\mu$ must have a non-zero basin of attraction. For this reason, one may introduce [7, 10, 18] a positive threshold κ, and demand the stronger inequalities

$$\gamma_i^\mu(t) \geqslant \kappa \tag{11}$$

to hold, rather than the inequalities (10), which are equivalent to the fixed point equations (7). The larger the threshold κ, the larger the basins of attractions can be expected to be [10, 18].

In order to solve equation (11) for the unknown weights w_{ij}, we consider it as far as its equality sign is concerned. Then (11) can be recast in the equivalent form

$$\sum_{j=1}^{N} w_{ij}(t)\xi_j^\mu - \theta_i = \kappa(2\xi_i^\mu - 1) \qquad (i = 1, \ldots, N; \mu = 1, \ldots, p) \tag{12}$$

as may be checked by putting ξ_i^μ equal to 1 or 0. The pN equations (12) do not fix uniquely the N^2 weights w_{ij} as long as $p < N$, the case we consider throughout this article. The storage capacity α, defined as $\alpha := p/N$, of a neural network is maximally equal to one for networks described by equations (12).

Various types of networks. It is our aim to take into account specific aspects of the connectivity of a biological network. In a biological neural network a neuron does not excite or inhibit itself, i.e. for all t we have for the self-interactions (or self-connections)

$$w_{ii}(t) = 0 \qquad (i = 1, \ldots, N). \tag{13}$$

Moreover, a biological network will, in general, be partially connected: each neuron will have some neighbourhood outside which there are no connections, i.e.

$$w_{ij}(t) = 0 \tag{14}$$

for a given set of neuron pairs (i, j). We shall call a network in which a (finite) fraction of the weights vanish, a diluted network. Let M_0 be the number of pairs (i, j) for which $w_{ij}(t) = 0$. Then the dilution d of a network of N neurons is defined as

$$d := M_0/N^2. \tag{15}$$

Hence, the dilution d is a number between 0 and 1.

Let us slightly generalize the above by distinguishing in a learning process changing and non-changing connections $w_{ij}(t)$ instead of changing and vanishing connections. Let us consider, for a moment, one particular neuron i. Then one may define the index sets

$$V_i := \{j | w_{ij}(t) \neq w_{ij}(t_0)\} \qquad V_i^c := \{j | w_{ij}(t) = w_{ij}(t_0)\}. \tag{16}$$

Thus V_i contains the indices related to all connections of neuron i that, in a learning process, change in time, whereas its complement, V_i^c, contains the indices related to all non-changing connections. In particular V_i^c contains the index of neuron i itself ($w_{ii}(t) = w_{ii}(t_0) = 0$), the indices of neurons j which have no connections with neuron i ($w_{ij}(t) = w_{ij}(t_0) = 0$), and the indices of neurons j which have connections with fixed strengths with neuron i ($w_{ij}(t) = w_{ij}(t_0) \neq 0$). Thus, diluted networks are a subclass of networks with changing and non-changing connections. By specifying, via equation (16), which connections are absent, the network connectivity is completely defined. For later use, we introduce M, the number of pairs (i, j) for which $w_{ij}(t) = w_{ij}(t_0)$ is constant, but not necessarily equal to zero.

3. Learning prescriptions—Hebb rules

In this section we consider all mathematical realizations which are, in principle, compatible with Hebb's postulate. We argue that, in our view, only two of them, namely (20) and (21) are biologically plausible, in contrast to the realizations (22) and (23) used in the literature. In order to show this, let us consider a network with changing and non-changing connections, in which a learning process takes place with the purpose of storing a collection of p patterns $\boldsymbol{\xi}^\mu$. Let the weights at time t_n be given by $w_{ij}(t_n)$. After a learning step the new weights will be given in terms of the old weights by

$$w_{ij}(t_{n+1}) = \begin{cases} w_{ij}(t_n) + \Delta w_{ij}(t_n) & (j \in V_i) \\ w_{ij}(t_n) & (j \in V_i^c) \end{cases} \tag{17}$$

where $\Delta w_{ij}(t_n)$ is the increment at time t_n. A learning rule is a recipe for the change Δw_{ij} as a function of the states of the post-synaptic neuron i and the pre-synaptic neuron j when a pattern $(\xi_1, \ldots, \xi_N)$ is presented to the network. There are four possible states (ξ_i, ξ_j) that the post- and pre-synaptic neuron can have, namely (0, 0), (0, 1), (1, 0) and (1, 1), each of which may lead to one of the three possible changes for Δw_{ij}: positive, negative or zero. Hence, in principle there are $3^4 = 81$ possible learning rules

$$\Delta w_{ij} : (\xi_i, \xi_j) \longmapsto \Delta w_{ij}(\xi_i, \xi_j). \tag{18}$$

It is biologically improbable that connections will always grow or will always decrease. Therefore, we exclude learning rules for which $\Delta w_{ij}(\xi_i, \xi_j)$, for all four states (ξ_i, ξ_j), are either always positive, or always negative (reason of rejection *a* of table 1). Moreover, in our opinion, it is biologically probable that a connection between a pre-synaptic neuron j and a post-synaptic neuron i does not change if the neuron j does not contribute to the post-synaptic potential of neuron i, i.e. if $\xi_j = 0$. Therefore, we exclude learning rules for which $\Delta w_{ij}(\xi_i, \xi_j = 0) \neq 0$ with $\xi_i = 0, 1$ (reason of rejection *b* of table 1).

Excluding these improbable learning rules, we are left with no more than two learning rules, as may be verified by a simple inspection of table 1. One of these corresponds to the assignments

$$\begin{aligned} (0,0) &\longmapsto \Delta w_{ij} = 0 \qquad & (0,1) &\longmapsto \Delta w_{ij} < 0 \\ (1,0) &\longmapsto \Delta w_{ij} = 0 \qquad & (1,1) &\longmapsto \Delta w_{ij} > 0 \end{aligned} \tag{19}$$

(column H in table 1), which can be expressed compactly by the formula

$$\Delta w_{ij} = \epsilon_{ij}(2\xi_i - 1)\xi_j \tag{20}$$

where the ϵ_{ij}, here and elsewhere in this paper, are positive numbers. Similarly, the other one can be expressed by the formula

$$\Delta w_{ij} = -\epsilon_{ij}(2\xi_i - 1)\xi_j \tag{21}$$

Table 1. The 81 possible ways in which w_{ij} may change as a function of the activities of the post-synaptic neuron i and the pre-synaptic neuron j can be read off from the 81 columns of the table. Each row may have up arrows (↑), down arrows (↓) or zeros, indicating a strengthening, a weakening or no change of a synaptic connection. The biological reason to reject a column is indicated by the letter *a* or *b* immediately below the column. The reasons are *a*: there either is only strengthening or weakening of the synapse, *b*: there is a change of the synaptic strength if the pre-synaptic neuron j is inactive. From the table we can read off that 78 possibilities are excluded for reason *a* and/or *b*. The column with only zeros is excluded for obvious reasons. The two possibilities for the Hebb rule which we are left with are indicated by the symbols *H* and *A*: the first corresponds to what is called Hebbian learning, the second to what is called anti-Hebbian learning. If we do not reject a possibility for reason *b*, there are many more possible Hebbian rules. The possibility indicated by *G* was used by Gardner [19]. The one preferred by physicists in their modelling of neural networks, has been indicated by the symbol *P*.

(i,j)						*G*		*H*									*P*								
$(0,0)$	0	0	0	0	0	0	0	0	0	↑	↑	↑	↑	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓	↓
$(0,1)$	↑	↑	↑	0	0	0	↓	↓	↓	↑	↑	↑	0	0	0	↓	↓	↓	↑	↑	↑	0	0	0	↓	↓	↓
$(1,0)$	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓
$(1,1)$	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	a	*a*		*a*	*a*					*a*	*a*		*a*	*a*													
	b		*b*	*b*		*b*	*b*		*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*
(i,j)		*A*																									
$(0,0)$	0	0	0	0	0	0	0	0	0	↑	↑	↑	↑	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓	↓
$(0,1)$	↑	↑	↑	0	0	0	↓	↓	↓	↑	↑	↑	0	0	0	↓	↓	↓	↑	↑	↑	0	0	0	↓	↓	↓
$(1,0)$	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓
$(1,1)$	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
					a	*a*		*a*	*a*													*a*	*a*		*a*	*a*	
	b		*b*	*b*		*b*	*b*		*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*
(i,j)																											
$(0,0)$	0	0	0	0	0	0	0	0	0	↑	↑	↑	↑	↑	↑	↑	↑	↑	↓	↓	↓	↓	↓	↓	↓	↓	↓
$(0,1)$	↑	↑	↑	0	0	0	↓	↓	↓	↑	↑	↑	0	0	0	↓	↓	↓	↑	↑	↑	0	0	0	↓	↓	↓
$(1,0)$	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓	↑	0	↓
$(1,1)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	a	*a*		*a*		*a*		*a*	*a*	*a*	*a*		*a*	*a*								*a*	*a*		*a*	*a*	
	b		*b*	*b*		*b*	*b*		*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*	*b*

(column *A* in table 1).

Learning can be classified as Hebbian or anti-Hebbian. Hebbian learning is characterized by the fact that, if both neurons i and j are active, Δw_{ij} is positive, whereas for anti-Hebbian learning Δw_{ij} is negative. So, the two remaining learning rules (20) and (21) are Hebbian and anti-Hebbian, respectively. The learning rules (20) and (21) have, to the best of our knowledge, not been used, as yet, in mathematical or physical studies that tried to model biological neural systems (see, e.g. [11, 20]).

If we allow for the possibility that $\Delta w_{ij} \neq 0$ if the pre-synaptic neuron j is inactive ($\xi_j = 0$), there are many extra possible mappings (18), of which we mention the two most often encountered in the literature

$$\Delta w_{ij} = \epsilon_{ij}\xi_i(2\xi_j - 1) \tag{22}$$

$$\Delta w_{ij} = \epsilon_{ij}(2\xi_i - 1)(2\xi_j - 1). \tag{23}$$

The learning rule (22) was used, e.g. by Gardner [19], in studying the retrieval properties of a neural network with an asymmetric learning rule (column *G* in table 1). The learning rule (23) is the one most often used by physicists [20, 21] in their modelling of neural networks (row *P* in table 1).

Finally, let us compare the four learning rules (20)–(23) after one learning step of one pattern $\boldsymbol{\xi}$. Let us suppose that a pattern $\boldsymbol{\xi}$ is not yet learned at time t_0 so that, in view of

(11), the quantity $\gamma_i(t_0)$ is negative. In order to store a pattern, γ_i should be positive. Upon substitution of the Hebbian or symmetric learning rules (20) or (23) into (8) we find

$$\gamma_i(t_1) = \gamma_i(t_0) + \sum_{j \in V_i} \epsilon_{ij}\xi_j \tag{24}$$

for the anti-Hebbian learning rule (21) we get

$$\gamma_i(t_1) = \gamma_i(t_0) - \sum_{j \in V_i} \epsilon_{ij}\xi_j \tag{25}$$

whereas for the asymmetric learning rule (22) we obtain

$$\gamma_i(t_1) = \gamma_i(t_0) + \xi_i \sum_{j \in V_i} \epsilon_{ij}\xi_j \tag{26}$$

where t_0 is the initial time and t_1 is the time after one learning step. By a suitable choice for ϵ_{ij} it can *always* be achieved that $\gamma_i(t_1)$ is positive in the cases of the Hebbian and symmetric learning rules (20) and (23), whatever the values of ξ_i and ξ_j, as follows from (24). This can *never* be achieved in the case of the anti-Hebbian learning rule (21), as is seen from (25). Finally, in the case $\xi_i = 0$, this can *never* be achieved for the asymmetric learning rule (22), as can be read off from (26). These simple arguments show that the Hebbian and symmetric learning rules (20) and (23)—but not the anti-Hebbian and asymmetric learning rules (21) and (22)—are, in principle, suitable for storage of patterns.

In the next section we will show that the requirement that synaptic changes take place in an energetically economic way leads to a learning rule which, depending on the state of the post-synaptic neuron i, is of the Hebbian or anti-Hebbian form (20) or (21). Hence, the naive approach of this section, which leads to the two forms (20) and (21), is consistent with an approach which is based on a physical principle.

4. Energy saving learning rule

In the literature, Hebb rules for the change of the synaptic connections have been derived in various manners, many of which essentially correspond to the determination of an extremum of some 'Lyapunov' or 'cost function', also called 'energy function'

$$H(t) = -\tfrac{1}{2} \sum_{i,j=1}^{N} J_{ij}(t)s_i(t)s_j(t). \tag{27}$$

If $J_{ij} = J_{ji}$, equation (27) is the central equation of the Hopfield model [21]. In the case of an Ising system of atoms with spin, an equation of the form (27) corresponds to the actual physical energy of the spin-system.

For a system of neurons, however, an energy function of the form (27) is an ad hoc postulate. It is not derived from or suggested by some underlying biological, biochemical or biophysical principle. In other words, the function (27), is, *a priori*, totally unrelated to the actual energy of the neural system. Consequently, a 'derivation' leading to a Hebb rule based on a function of the type (27), (see e.g. [14]), is just as ad hoc as the postulate underlying it.

In this section we will show that the Hebb rule (20) and its anti-Hebbian counterpart (21) can be found by postulating that the (*biochemical*) *energy* needed to change the synapses—in order to store a new pattern $\boldsymbol{\xi}$—is minimal. We thus show that these particular Hebb rules—and only these ones—are consistent with a physical principle. The argument runs as follows.

The energy ΔE_{ij} to change the connection $w_{ij}(t_n)$ to $w_{ij}(t_{n+1})$ will be a differentiable function of the magnitude of the change $\Delta w_{ij}(t_n)$ occurring in (17)

$$\Delta E_{ij} = f_{ij}(\Delta w_{ij}). \tag{28}$$

If a synapse between the neurons i and j is not changed in a learning step there is no energy consumed. Hence, the energy change ΔE_{ij} vanishes if $\Delta w_{ij} = 0$, i.e.

$$f_{ij}(0) = 0. \tag{29}$$

Moreover, we assume that a change of a synapse, whether it be a strengthening or a weakening, can only be achieved by *adding* energy to the system. Thus, if $\Delta w_{ij} \neq 0$, we put,

$$f_{ij}(\Delta w_{ij}) > 0. \tag{30}$$

Equations (29) and (30) enable us to obtain a useful approximate expression for the energy change ΔE_{ij}. We first note that any differential function $f(x)$ can be written as a power series $f(x) = c^{(0)} + c^{(1)}x + c^{(2)}x^2 + \cdots$. Thus, we have for the function (28), up to terms quadratic in Δw_{ij},

$$f_{ij}(\Delta w_{ij}) = c_{ij}^{(0)} + c_{ij}^{(1)}\Delta w_{ij} + c_{ij}^{(2)}\Delta w_{ij}^2 \tag{31}$$

where, in view of (29) and (30) the coefficients have the properties

$$c_{ij}^{(0)} = 0 \qquad c_{ij}^{(1)} = 0 \qquad c_{ij}^{(2)} > 0. \tag{32}$$

Furthermore, we take

$$c_{ij}^{(2)} = c_i \tag{33}$$

which is equivalent to the supposition that a change of connections related to different synapses $j = 1, 2, \ldots, N$ of the same neuron i needs the same amount of energy. This assumption simplifies some of the formulae below; it is not essential in the sense that all conclusions remain unaltered if the simplification (33) is not used, see [23]. The total change ΔE in the nth learning step $w_{ij}(t_n) \longrightarrow w_{ij}(t_{n+1})$, where in principle all w_{ij} with $j \in V_i$ may change, is given by the sum of the individual changes,

$$\Delta E(\Delta w_{kl}) = \sum_{i=1}^{N} \sum_{j \in V_i} f_{ij}(\Delta w_{ij}) \tag{34}$$

or, inserting (31) with (32) and (33), by

$$\Delta E(w_{kl}(t_{n+1})) = \sum_{i=1}^{N} \sum_{j \in V_i} c_i (w_{ij}(t_{n+1}) - w_{ij}(t_n))^2. \tag{35}$$

The positive constants c_i are characteristic of neuron i.

Equation (35) will be our starting point for the derivation of the energy saving learning rule (42). It is the general form any expression must have that describes the energy needed to adapt the connection strengths w_{ij} as a function of their changes Δw_{ij}. We now will minimize the change in energy ΔE as a function of the new weights $w_{kl}(t_{n+1})$ under the constraint (12) using the Lagrange method. This was the reason to write ΔE in (35) as a function of the $w_{kl}(t_{n+1})$ rather than as a function of the $\Delta w_{kl} = w_{kl}(t_{n+1}) - w_{kl}(t_n)$, as was done in (34).

4.1. Storage of one pattern

Let us consider at the nth learning step, i.e. at time t_n, the storage of one pattern $\boldsymbol{\xi}$ in a network with connections given by $w_{ij}(t_n)$. In the case of a network with changing and non-changing weights as introduced in section 2, the expression for the change of energy is, up to second order in the changes of the synaptic weights, given by (35). Note that a minimization of the one condition (35) under the constraint induced by the fixed point equation (12) implies a minimization of the $N^2 - M$ changes $\Delta w_{ij}^2(t_n)$, since a sum of positive terms is minimal if and

only if each term is minimal; recall that M is the number of synapses with constant weights w_{ij}.

For the storage of one single pattern $\boldsymbol{\xi}$, one may rewrite the fixed point equations (12) in the form

$$g_i(w_{ij}(t_{n+1})) = 0 \qquad (i = 1, \ldots, N) \tag{36}$$

where

$$g_i(w_{ij}(t_{n+1})) = \kappa(2\xi_i - 1) - \sum_{j \in V_i^c} w_{ij}(t_n)\xi_j - \sum_{j \in V_i} w_{ij}(t_{n+1})\xi_j + \theta_i. \tag{37}$$

The method of Lagrange multipliers tells that one finds the extrema of (35) subject to the auxiliary conditions (36) from the $N^2 - M$ equations

$$\frac{\partial \Delta E}{\partial w_{ij}(t_{n+1})} + \sum_{k=1}^{N} \lambda_k \frac{\partial g_k}{\partial w_{ij}(t_{n+1})} = 0 \qquad (i = 1, \ldots, N;\ j \in V_i) \tag{38}$$

Upon substitution of (35) and (37) into this expression, we find the $N^2 - M$ relations

$$w_{ij}(t_{n+1}) = w_{ij}(t_n) + \frac{1}{2c_i}\lambda_i \xi_j \qquad (i = 1, \ldots, N;\ j \in V_i). \tag{39}$$

In the method of Lagrange multipliers the number of constraints equals the number of Lagrange multipliers λ_i. Hence, there are N Lagrange multipliers. Since the N multipliers λ_i are unequal to zero, it follows from the $N^2 - M$ equations (39) that $N^2 - M \geqslant N$, or $M \leqslant N^2 - N$. We now have obtained the $N + N^2 - M$ equations (36) and (39) for the $N + N^2 - M$ unknowns λ_i and $w_{ij}(t_{n+1})$.

The structure of these equations happens to be such that an explicit expression for the λ_i can be found, and thereupon, an explicit expression for the $w_{ij}(t_{n+1})$ can be obtained. The procedure is as follows.

Eliminating the $w_{ij}(t_{n+1})$ from (36) with the help of (39), leads to

$$\lambda_i = \frac{2c_i}{\sum_{k \in V_i} \xi_k}[\kappa - \gamma_i(t_n)](2\xi_i - 1) \tag{40}$$

where we used the property $(\xi_j)^2 = \xi_j$. Substituting this expression for λ_i into (39) yields

$$w_{ij}(t_{n+1}) = w_{ij}(t_n) + \frac{1}{\sum_{k \in V_i} \xi_k}[\kappa - \gamma_i(t_n)](2\xi_i - 1)\xi_j \qquad (j \in V_i) \tag{41}$$

or, equivalently (see equation (17)),

$$\Delta w_{ij}(t_n) = \frac{1}{\sum_{k \in V_i} \xi_k}[\kappa - \gamma_i(t_n)](2\xi_i - 1)\xi_j \qquad (j \in V_i) \tag{42}$$

where κ is the positive parameter (11) related to the basins of attraction, and where the γ_i $(i = 1, \ldots, N)$ are the stability coefficients given by (8). We will refer to (42) by the name of *non-local energy saving learning rule*, since the denominator of (42) depends on the input from all neurons k that are connected via changing connections to neuron i. The factor between square brackets

$$\kappa - \gamma_i(t_n) = \kappa - (h_i(t_n) - \theta_i)(2\xi_i - 1) \tag{43}$$

depends solely upon the temporal and environmental state of the post-synaptic neuron i, that is, on its post-synaptic potential h_i at time t_n of the nth learning step, its thresholds θ_i, its activity ξ_i and a parameter κ. The factor (43) can be positive or negative. Therefore, the learning rule (42), (43) derived here from the assumption of minimal energy change per learning step, happens

to coincide with the particular Hebbian learning rule (20) and its anti-Hebbian counterpart (21) found in section 3 on purely intuitive grounds, grounds which were related to biological plausibility.

We thus have shown that if biological neurons would adapt their connections according to the non-local energy saving learning rule (42), this adaptation would be such that the network would fulfil the fixed point equation (12) for a pattern $\boldsymbol{\xi}$. Moreover, the learning rule (42) guarantees that the energy needed to rebuild a neural network with connections $w_{ij}(t_n)$ to a network with connections $w_{ij}(t_{n+1})$ is minimal.

We conclude this section with some remarks. The energy saving learning rule is only applicable in those situations in which the denominator is unequal to zero. This can be translated into a restriction on the ξ_k, $k \in V_i$. It follows that with an decreasing number of adaptable connections there is an increasing number of patterns that cannot be stored with the help of the non-local energy saving learning rule. This effect will be absent when the local energy saving learning rule is used (see section 6).

When we repeat the derivation of (42) in the spin representation with time-dependent thresholds as given by (5), we find again (42) with ξ replaced by $(s+1)/2$, i.e.

$$\Delta J_{ij} \propto s_i(s_j + 1) \tag{44}$$

as could be expected. If, however, the derivation of (42) is repeated in the spin representation with T_i taken to be a constant, as is usually done in the spin representation, one finds a result which differs from (44), namely

$$\Delta J_{ij} \propto s_i s_j. \tag{45}$$

This is the biologically less relevant result commonly encountered in the physical literature, as noticed already in section 3 (see equation (23)).

4.2. Storage of p patterns

In the previous section we saw that storage of one pattern $\boldsymbol{\xi}$ can be achieved via a synaptic change Δw_{ij} given by (42). Hence, storage of p patterns $\boldsymbol{\xi}^\mu$ ($\mu = 1, \ldots, p$) might be accomplished by repeated application of the learning rule (42). Let us therefore consider the following learning process. In a first interval of time, $[t_0, t_1)$, a first pattern $\boldsymbol{\xi}^1$ is stored via the change $\Delta w_{ij}(t_0)$, leading to the connections $w_{ij}(t_1) = w_{ij}(t_0) + \Delta w_{ij}(t_0)$, $j \in V_i$. Next, in the interval $[t_1, t_2)$, pattern $\boldsymbol{\xi}^2$ is stored, etc. Finally, pattern $\boldsymbol{\xi}^p$ is stored. We call this sequence of storage of p patterns a learning cycle.

The energy saving learning rule is a storage prescription for a new pattern, which does not take into account, however, any constraint that would guarantee that a previously stored patterns remain stored. Thus it may occur that storage of a new pattern will perturb, partially or totally, the storage of an older pattern.

In section 5, on maximal learning efficiency, we will determine a learning rule which does guarantee that new patterns are stored without wiping out previously stored patterns. However, this learning rule will turn out to be biologically unacceptable. We therefore proceed with the learning rule derived above. We shall derive, along the lines of reasoning of Diederich and Opper [7], but for diluted networks, an expression for the weights w_{ij} of the synaptic connections after infinitely many learning cycles. It will turn out that, in the end, previously stored patterns are not forgotten.

As follows from equation (17), the connections after R learning cycles are given by

$$w_{ij}(t_{Rp}) = w_{ij}(t_0) + \sum_{m=1}^{R}\sum_{\mu=1}^{p} \Delta w_{ij}(t_{(m-1)p+\mu-1}) \qquad (j \in V_i) \tag{46}$$

with t_{Rp} the time after R learning cycles of p patterns.

Substituting (42) into (46) we find

$$w_{ij}(t_{Rp}) = w_{ij}(t_0) + N^{-1}\sum_{\mu=1}^{p} F_i^{\mu}(t_{(R-1)p+\mu-1})\xi_j^{\mu} \qquad (j \in V_i) \tag{47}$$

where

$$F_i^{\mu}(t_{(R-1)p+\mu-1}) = \sum_{m=1}^{R}\left[\kappa(2\xi_i^{\mu}-1) - \left(\sum_{k\in V_i^c} w_{ik}(t_0)\xi_k^{\mu} + \sum_{k\in V_i} w_{ik}(t_{(m-1)p+\mu-1})\xi_k^{\mu} - \theta_i\right)\right] \times\left(N^{-1}\sum_{k\in V_i}\xi_k^{\mu}\right)^{-1} \tag{48}$$

is the effect on w_{ij} of pattern $\boldsymbol{\xi}^{\mu}$ after R learning cycles have been completed. From (48) it follows that

$$\left(N^{-1}\sum_{k\in V_i}\xi_k^{\mu}\right)[F_i^{\mu}(t_{(R-1)p+\mu-1}) - F_i^{\mu}(t_{(R-2)p+\mu-1})] = \kappa(2\xi_i^{\mu}-1) - \left(\sum_{k\in V_i^c} w_{ik}(t_0)\xi_k^{\mu} + \sum_{k\in V_i} w_{ik}(t_{(R-1)p+\mu-1})\xi_k^{\mu} - \theta_i\right). \tag{49}$$

In the Rth learning cycle, at time $t_{(R-1)p+\nu-1}$, only the patterns $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^{\nu-1}$ have changed the weights of the network. Hence, the F_i^{ν} with $\nu < \mu$ have new values at time $t_{(R-1)p+\mu-1}$, whereas the F_i^{ν} with $\nu \geqslant \mu$ are still identical to their values in the previous learning cycle, i.e. are equal to the values at time $t_{(R-2)p+\mu-1}$. Thus, with the help of (47), the weights in the right-hand side of (49) can be expressed as follows in terms of the F_i^{μ}:

$$w_{ik}(t_{(R-1)p+\mu-1}) = w_{ik}(t_0) + N^{-1}\sum_{\nu<\mu} F_i^{\nu}(t_{(R-1)p+\mu-1})\xi_k^{\nu} + N^{-1}\sum_{\nu\geqslant\mu} F_i^{\nu}(t_{(R-2)p+\mu-1})\xi_k^{\nu}. \tag{50}$$

Eliminating $w_{ik}(t_{(R-1)p+\mu-1})$ from (49) with the help of (50) yields

$$N^{-1}\sum_{k\in V_i}\sum_{\nu\leqslant\mu} F_i^{\nu}(t_{(R-1)p+\mu-1})\xi_k^{\nu}\xi_k^{\mu} = -N^{-1}\sum_{k\in V_i}\sum_{\nu>\mu} F_i^{\nu}(t_{(R-2)p+\mu-1})\xi_k^{\nu}\xi_k^{\mu} + [\kappa - \gamma_i^{\mu}(t_0)](2\xi_i^{\mu}-1). \tag{51}$$

This system of linear equations can be solved for F_i^{μ} using the Gauss–Seidel iterative method. We first rewrite (51) in matrix notation. Next, we introduce a $p \times p$ matrix C_i, the matrix elements of which are given by

$$C_i^{\mu\nu} := N^{-1}\sum_{k\in V_i}\xi_k^{\mu}\xi_k^{\nu}. \tag{52}$$

We might call this matrix the 'reduced correlation matrix', since it correlates ξ_k^{μ} and ξ_k^{ν} while taking into account, via V_i, the connectivity of the network. The reduced correlation matrix is closely related to the usual correlation matrix if V_i contains all neuron indices. We proceed by decomposing this matrix C_i into matrices L_i and U_i in such a way that $C_i = L_i + U_i$. The matrix L_i is a matrix with only non-zero matrix elements on and below the diagonal and U_i is a matrix with only non-zero matrix elements above the diagonal. We also introduce the vectors $\boldsymbol{F}_i(R) := (F_i^1(t_{(R-1)p+1-1)}), \ldots, F_i^p(t_{(R-1)p+p-1}))$ and $\boldsymbol{G}_i := ([\kappa - \gamma_i^1(t_0)](2\xi_i^1 - 1), \ldots, [\kappa - \gamma_i^p(t_0)](2\xi_i^p - 1))$. Finally, we shall denote a $p \times p$ unit matrix as I. We thus can rewrite (51) in the form

$$L_i \cdot \boldsymbol{F}_i(R) = -U_i \cdot \boldsymbol{F}_i(R-1) + \boldsymbol{G}_i. \tag{53}$$

By iteratively solving this equation for $\boldsymbol{F}_i(R)$, we find

$$\boldsymbol{F}_i(R) = [-L_i^{-1} \cdot U_i]^{R-1} \cdot \boldsymbol{F}_i(1) + L_i^{-1} \cdot [I - L_i^{-1} \cdot U_i + \cdots + (-L_i^{-1} \cdot U_i)^{R-2}] \cdot \boldsymbol{G}_i. \quad (54)$$

The symmetric matrix C_i, as defined in (52), is positive definite and symmetric. It then can be shown that the matrix $-L_i^{-1} \cdot U_i$ has eigenvalues smaller than one [22]. As a consequence, we have

$$\lim_{R\to\infty} [-L_i^{-1} \cdot U_i]^{R-1} = 0 \quad (55)$$

and it follows that, in the limit $R \to \infty$, (54) converges to

$$\begin{aligned} \boldsymbol{F}_i(\infty) &= L_i^{-1} \cdot [I - (-L_i^{-1} \cdot U_i)]^{-1} \cdot \boldsymbol{G}_i \\ &= C_i^{-1} \cdot \boldsymbol{G}_i \end{aligned} \quad (56)$$

where $\boldsymbol{F}_i(\infty) = \lim_{R\to\infty} \boldsymbol{F}_i(R)$. Substitution of (56) in (47) and restoring the old notation, yields, for $R \to \infty$

$$w_{ij}(t_\infty) = \begin{cases} w_{ij}(t_0) + N^{-1} \displaystyle\sum_{\mu,\nu=1}^{p} [\kappa - \gamma_i^\mu(t_0)](2\xi_i^\mu - 1)(C_i^{-1})^{\mu\nu}\xi_j^\nu & (j \in V_i) \\ w_{ij}(t_0) & (j \in V_i^c) \end{cases} \quad (57)$$

where $(C_i^{-1})^{\mu\nu}$ is the inverse of the matrix (52).

By substituting (57) into (12) it can directly be verified that the weights (57) fulfil (12) for all μ ($\mu = 1, \ldots, p$). For $p = 1$ this was to be expected, since the learning rule (42) was constructed that way. For $p > 1$ one could, for the same reason, expect that (12) would be verified by (57) for the final pattern of the learning cycle, $\boldsymbol{\xi}^p$. It is less transparent, however, that (57) satisfies (12) for all patterns $\boldsymbol{\xi}^\mu$.

The result (57) is exact for networks with a number of vanishing connections running from $M_0 = 0$ to $M_0 = N^2 - Np$, i.e., valid for dilution 0 to $d = 1 - \alpha$, where $\alpha = p/N$. The analogous calculation performed by Diederich and Opper for networks with empty V_i^c, so that V_i contains all indices, yields a result that coincides with the result obtained via the usual pseudo-inverse solution [5, 6] of equation (12). Hence, the following question may now arise. Can we solve the equation (12) for a neural network where V_i^c is not empty and, consequently, the method of the pseudo-inverse in its standard form is not applicable? The answer to this question is affirmative. In appendix B we modify the method of the pseudo-inverse so as to be applicable to systems with changing and non-changing interactions. Solving equation (12) for networks with changing and non-changing connections via what we have called the modified method of the pseudo-inverse, one indeed obtains (57), as we also prove in the appendix.

Thus we have shown that the solution that corresponds to the stepwise energetically most economic way to realize storage of patterns in a partially connected network, turns out to be identical to the one obtained via a—modified—version of the well known mathematical method of the pseudo-inverse applied to the fixed point equation (12). In other words, the non-local energy saving learning rule (42) leads to the solution of the fixed point equation (12), obtained via the modified method of the pseudo-inverse, which is based, in turn, on the reduced correlation matrix.

We conclude this section with a few remarks. In general, the inverse of the matrix $C_i^{\mu\nu}$ cannot easily be found analytically. However, in the non-biological case that none of the weights is kept constant, all index sets V_i^c are empty. As a consequence one may use, for large N and low storage capacity $\alpha := p/N$, the approximations

$$N^{-1} \sum_{j=1}^{N} \xi_j^\mu = a \quad (58)$$

$$N^{-1}\sum_{j=1}^{N}\xi_j^\mu\xi_j^\nu = a^2 \qquad (\mu \neq \nu). \tag{59}$$

Substitution of (58) and (59) into (52), where now V_i is the set of all indices, yields

$$C_i^{\mu\nu} = a(1-a)\delta_{\mu\nu} + a^2. \tag{60}$$

For the inverse of $C_i^{\mu\nu}$ we thus obtain from (60) the simple analytical expression

$$(C_i^{-1})^{\mu\nu} = \frac{1}{a(1-a)}\left[\delta_{\mu\nu} - \frac{a}{ap-a+1}\right]. \tag{61}$$

Using (61) in (57), leads to

$$w_{ij}(t_\infty) = w_{ij}(t_0) - \frac{1}{Na(1-a)}\frac{a}{ap-a+1}\sum_{\mu,\nu=1}^{p}[\kappa - \gamma_i^\mu(t_0)](2\xi_i^\mu - 1)\xi_j^\nu$$
$$+\frac{1}{Na(1-a)}\sum_{\mu=1}^{p}[\kappa - \gamma_i^\mu(t_0)](2\xi_i^\mu - 1)\xi_j^\mu \qquad (i, j = 1, \ldots, N). \tag{62}$$

Equation (62) is an explicit expression for the weights w_{ij} of a (non-biological) network in which all the weights, including the self-interactions w_{ii}, are present.

Kanter and Sompolinsky used the result (57) in case $i \neq j$ for a fully connected network without self-interactions [9]. Their ad hoc assumption that the self-interactions w_{ii} can be put equal to zero, turns out to be justified in view of our exact result in (57) with $w_{ii}(t_0) = 0$.

5. A learning rule with maximal learning efficiency

In the preceding section learning of a collection of patterns was achieved by repeated application of the non-local energy saving learning rule. This learning rule was not constructed in such a way that conservation of storage of old patterns was automatically guaranteed when a new pattern was stored. We now address the question of whether and how storage of a new pattern $\boldsymbol{\xi}^{p+1}$ can be achieved without disturbing the storage of the old patterns $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^p$. We shall refer to this type of learning as maximally efficient learning.

Linkevich [8] treated this problem on the basis of a mathematical model, in which suppositions are made which cannot be true in a biological neural network. Firstly, he treated the thresholds $T_i(t)$, equation (5), as a vanishing constant. Moreover, his network has symmetric connections $w_{ij}(t) = w_{ji}(t)$, whereas a biological network has non-symmetric connections $w_{ij}(t) \neq w_{ji}(t)$. Finally, his network is fully connected, i.e. all $w_{ij}(t) \neq 0$.

We may improve and generalize the reasoning of Linkevich to obtain a maximally efficient learning rule for a partially connected network with non-symmetric connections. The calculations only hold for networks in which the thresholds are equal to the stability coefficients κ, i.e. $\theta_i = \kappa$, for all i, and in the case where the initial connections are equal to zero, $w_{ij}(t_0) = 0$ for all i and j. As a final result we arrive, in this particular case, at the following rule for learning with maximal learning efficiency (see appendix A)

$$\Delta w_{ij}(t) = \frac{[\kappa - \gamma_i^{p+1}(t)][\kappa - \gamma_j^{p+1}(t)](2\xi_i^{p+1} - 1)(2\xi_j^{p+1} - 1)}{\sum_{l\in V_i}[\kappa - \gamma_l^{p+1}(t)]\xi_l^{p+1}} \qquad (j \in V_i). \tag{63}$$

From (63) we immediately see that, in general, Δw_{ij} is not symmetric in i and j. However, for a network in which all connections may change we find that Δw_{ij} is symmetric in i and j, in accordance with the result of Linkevich. Note that the i-dependent factors in the numerators of (63) and (42) are identical, which reflects the fact that the new pattern $\boldsymbol{\xi}^{p+1}$ has to obey

the fixed point equation, both in the cases of 'stepwise minimal change in energy' (42) and of 'stepwise maximal efficient learning' (63).

The learning rule with maximal learning efficiency (63) is of the form (23), a form which we have rejected, in section 3, on biological grounds. We therefore shall not pursue any further the analysis of the learning rule with maximal learning efficiency in the remainder of this paper.

6. Locality of learning rules

Up to now we have not mentioned an important limitation of a biological learning rule. The mathematical learning rule to change a weight of a network can, in principle, be local or non-local. The second possibility must be excluded in case a weight is associated with a synapse: there is no biological construction available in the brain to tell a specific synapse how and when to change as a function of properties of neurons with which it has no direct contact. The modifications must result from the *local* situation, i.e. limited to the situation spatially 'close enough' to the synapse in question, and within a 'brief span' of time. Thus, a change Δw_{ij} may depend only on variables local, in space and time, to the neurons i and j. The local variables available at the synapse between neurons i and j are the activities ξ_i and ξ_j, the post-synaptic potentials h_i and h_j, and the thresholds θ_i and θ_j. Hence, the factors ϵ_{ij} occurring in Hebb rules should depend on these variables only

$$\epsilon_{ij} = \epsilon_{ij}(\xi_i, h_i, \theta_i, \xi_j, h_j, \theta_j). \tag{64}$$

The energy saving learning rule (42) for Δw_{ij} guarantees, after repeated application, storage of patterns in a way which is energetically efficient. The factor between square brackets in the non-local learning rule (42) fulfils the criterion of locality. However, the learning rule as a whole is not a local learning rule because of the factor,

$$1 \Big/ \sum_{k \in V_i} \xi_k \tag{65}$$

which depends, because of the sum over k restricted to V_i, equation (16), on the network connectivity, and hence, not on properties related to neurons i and j only. If we approximate (65) by some constant, η_i say, we do obtain a learning rule that is local,

$$\Delta w_{ij}(t_n) = \eta_i[\kappa - (h_i(t_n) - \theta_i)(2\xi_i - 1)](2\xi_i - 1)\xi_j. \tag{66}$$

We shall refer to (66) as the *local energy saving learning rule*. The better η_i approximates a value dictated by (65), the better this local learning rule will be with respect to its energetic efficiency.

At this point it is important to note that the proof of convergence of section 4.2 can be generalized, replacing everywhere the factor (65) by the constant positive factor η_i. As a final result (57) is found again, provided certain restrictions on η_i are satisfied. It then can be proved [23] that the local, biologically realizable energy saving learning rule yields the same final values $w_{ij}(t_\infty)$ as the non-local energy saving learning rule.

As noticed in section 1, the constant η_i is a neuron property, the determination of which is outside the scope of this paper: we then would have to determine the coefficients c_{ij} in the expression for f_{ij} (31) explicitly.

A reasonable approximation for η_i can easily be obtained for a fully connected network where all connections may change in time. For such a network we have the approximation (58) for the denominator of (65), which implies

$$\eta_i \approx (Na)^{-1} \qquad \text{for all} \quad i = 1, \ldots, N. \tag{67}$$

We will use this approximation in the following section where we consider a biological network.

7. Local versus non-local learning

In this section, we will study numerically, for a biological network with dilution d, the local energy saving learning rule (66) as a competitor of the non-local learning rule (42). For η_i we take, quite arbitrarily, the constant (67). We could as well have taken $1/N$ or $1/(N(1-d))$: the essentials of the behaviour of the numerical results are not very sensitive for the precise values of the η_i.

In order to judge the functioning of a recurrent network with respect to its ability to store an arbitrary collection of p patterns $\boldsymbol{\xi}^\mu$ ($\mu = 1, \ldots, p$), we take L sets of such collections, and label them by $\boldsymbol{\xi}^{\mu,m}$ ($m = 1, \ldots, L$), i.e. $\boldsymbol{\xi}^{\mu,m}$ is pattern μ of set m. The performance of the network with respect to the patterns from the mth set may be characterized by the Np stability coefficients $\gamma_i^{\mu,m}$ ($i = 1, \ldots, N$; $\mu = 1, \ldots, p$) defined in equation (8). The stability coefficients $\gamma_i^{\mu,m}$ should be positive (see equation (10)). Moreover, we have normalized in such a way that the γ should be close to one. Hence, the more $\gamma_i^{\mu,m}$ we find with values around one, the better the network will perform.

We first define for the particular set m of p patterns the quantity:

$$\gamma^m = \min_{i=1,\ldots,N} \{\gamma_i^{1,m}, \ldots, \gamma_i^{p,m}\}. \tag{68}$$

Hence, γ^m is the minimal value of all stability coefficients for a particular set m of p patterns. A network does not function if γ^m is negative, and functions better and better when γ^m becomes closer to one (with the normalization $\kappa = 1$). To find a number that characterizes the network performance for an arbitrary set of p patterns, we average the minimal values γ^m over L arbitrarily chosen sets,

$$\gamma = \frac{1}{L} \sum_{m=1}^{L} \gamma^m. \tag{69}$$

Hence, γ is the average with respect to the L sets of p patterns ξ^μ. We therefore will refer to γ as the average performance of the network. Similarly, we define the average energy change ΔE

$$\Delta E = \frac{1}{L} \sum_{m=1}^{L} \Delta E^m \tag{70}$$

where ΔE^m is the change of energy in one learning step of the mth set of patterns. Furthermore, we define the average energy change per synapse Δe, as

$$\Delta e = \Delta E/(N^2 - M) \tag{71}$$

where M is the number of non-changing synapses. We will also study the performance of neural networks with varying dilution by considering the distribution of the stability coefficients $\gamma_i^{\mu,m}$. By studying numerically the quantities γ and Δe and the distribution of the stability coefficients $\gamma_i^{\mu,m}$, we can judge the power of the (exact) non-local energy saving learning rule (42) compared to the (biologically feasible) local energy saving learning rule (66), (67).

7.1. Storage of one pattern

Performance. The non-local energy saving learning rule (42) and its local approximation (66), (67) are used to store one pattern $\boldsymbol{\xi}$. In order to compare the quality of the two learning rules we have plotted in figure 1 the average performance γ versus the dilution d of the network for both learning rules. We see that the non-local learning rule stores a new pattern such that $\gamma = 1$, as could be expected since it has been designed that way. Moreover, we see that both

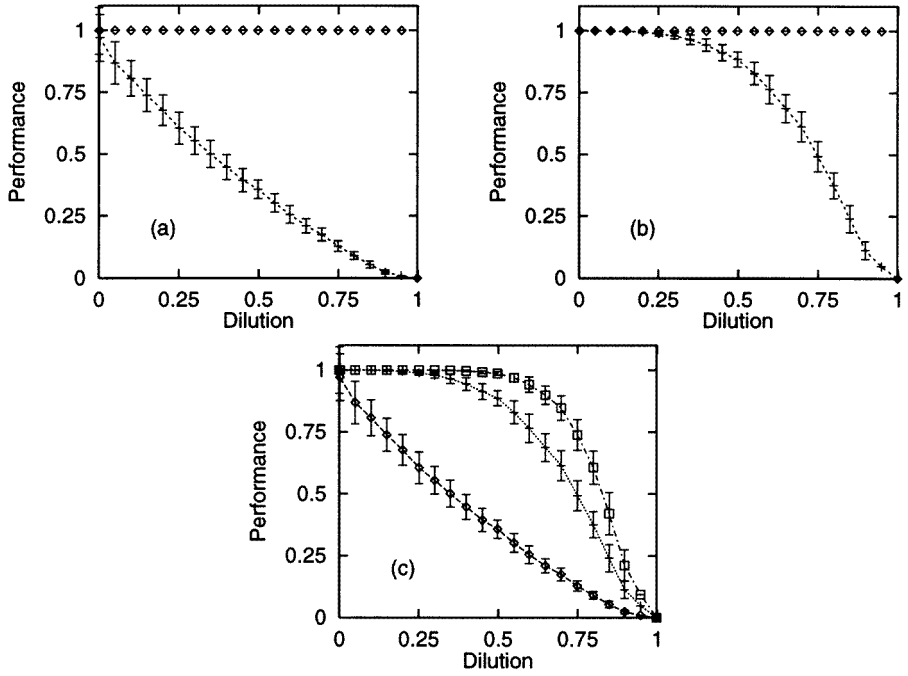

Figure 1. The average performance, γ, of a network of 512 neurons as a function of its dilution d. Dilution $d = 0$ means that the network is fully interconnected ($w_{ij} \neq 0$ for all i and j), dilution $d = 1$ means that there are no connections any more ($w_{ij} = 0$ for all i and j). The one pattern ξ is chosen arbitrarily, but such that the mean activity $a = 0.2$. The computations have been averaged over 100 different ξ. The error bars give the standard deviation of the averaged stability coefficients γ_i ($i = 1, \ldots, N$). The calculations are performed starting from a tabula rasa for the weights ($w_{ij}(t_0) = 0$) and vanishing thresholds ($\theta_i = 0$). (*a*), (*b*) In the first two figures, a comparison between the non-local energy saving learning rule (42) (upper curves) and the local energy saving learning rule (66) (lower curves) after it has been applied (*a*) one, and (*b*) five times. (*c*) A comparison of the local energy saving learning rule (66) after it has been applied one (lower curve), five and ten (upper curve) times.

the non-local and the local learning rules lead to positive values of γ, and, hence, lead to storage of the pattern ξ. The non-local learning rule, however, leads at once to $\gamma = 1$, whereas the local learning rule converges to $\gamma = 1$ only after repeated application. Hence, basins of attractions of the local learning rule are smaller initially (see figure 1).

Use of energy. Furthermore, we consider the average energy change per synapse Δe (71) for the non-local and local learning rules as a function of the number of synapses in a network of a fixed number of neurons. In the case of a single application of an energy saving learning rule, it turns out that for the non-local learning rule Δe increases as the number of synapses decreases, while Δe is constant in case of the local learning rule. This favourable situation of remaining constant apparently is an unexpected positive effect of the approximation made when going from a non-local energy saving learning rule to a local energy saving learning rule.

In the case of repeated application there is almost no energy effect for the non-local learning rule, and a slight effect for the local learning rule: the energy need per synapse grows with growing dilution (see figure 2).

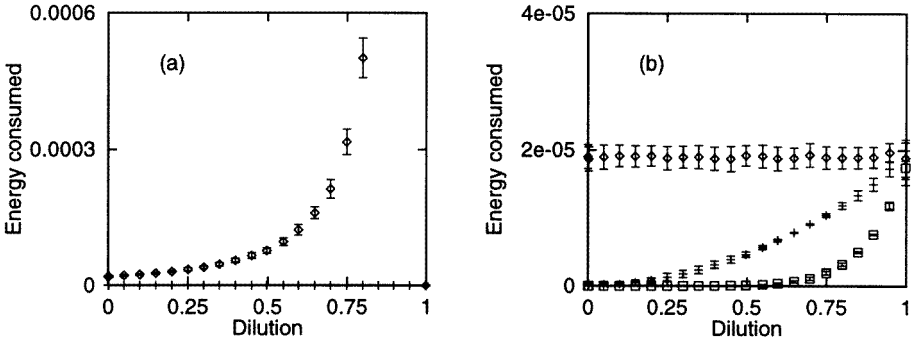

Figure 2. The average energy consumed per synapse Δe in one learning step, of a network of 512 neurons as a function of its dilution d. The one pattern $\boldsymbol{\xi}$ is chosen arbitrarily, but such that the mean activity $a = 0.2$. The computations have been averaged over 100 different $\boldsymbol{\xi}$. The error bars give the standard deviation of the averaged stability coefficients γ_i ($i = 1, \ldots, N$). The calculations are performed starting from a tabula rasa for the weights ($w_{ij}(t_0) = 0$) and vanishing thresholds ($\theta_i = 0$). (*a*) The average energy change per synapse Δe for the non-local energy saving learning rule after one (upper curve) and two learning steps (lower curve, coinciding with the horizontal axis). (*b*) The average energy change per synapse Δe for the local energy saving learning rule caused by the first (upper curve), second or fifth (lower curves) time that the local energy saving rule (66), (67) is used.

7.2. Storage of p patterns

Having studied numerically the storage of one pattern, we now turn to the storage of p patterns. As pointed out in section 4.2 this may be achieved through repeated application of the energy saving learning rule.

Storage of one pattern ($p = 1$) could be achieved in such a way that, by construction, all $\gamma_i^{\mu,m}$ ($\mu = 1$) were equal to one in case of the non-local learning rule: $\gamma_i^{1,m} = 1$ for all i and m. As a consequence, the local energy saving learning rule, which is an approximation to the non-local one, has the property that all $\gamma_i^{1,m}$ are 'not too far away' from the value $\kappa = 1$, i.e. they are positive. We recall that positivity of the stability coefficients $\gamma_i^{1,m}$ is a sufficient criterion for a network to store what should be stored (see figure 1).

When the energy saving learning rule is used to store more than one pattern, the positivity of all but the last stored pattern is not guaranteed. As noted before, we must allow for the fact that storage of a new pattern may spoil the storage of older patterns. Therefore, the requirement that the minimum of all $\gamma_i^{\mu,m}$ ($\mu = 1, \ldots, p$) should be positive is too strong. Forgetting this turns out to be an inevitable consequence of storing new patterns, at least in the beginning. By repeating the learning procedure for whole sequences of patterns more and more $\gamma_i^{\mu,m}$ become positive, suggesting that more and more patterns may be definitely stored.

In order to judge the performance of the network in the case of storage of more patterns, we now picture the distribution of the $\gamma_i^{\mu,m}$ over the real axis. Ideally, all $\gamma_i^{\mu,m}$ should be equal to $\kappa = 1$. In figure 3 the distribution has been plotted for both the non-local and local energy saving learning rule. As one observes from figure 3, some of the γ have values smaller than one (and even negative) whereas others have values larger than one. This is due to the fact that storing in set m a pattern $\boldsymbol{\xi}^\nu$, the $\gamma_i^{\mu,m}$ of the other patterns $\boldsymbol{\xi}^\mu$ ($\mu \neq \nu$) are not taken into account in the learning step and as a consequence can be enlarged or reduced in value. We have chosen to put the number of γ with values outside the plotted interval in the very first and the very last interval (see, e.g. figure 3(*e*)).

The general conclusion is that the local energy-saving learning rule, although in principle

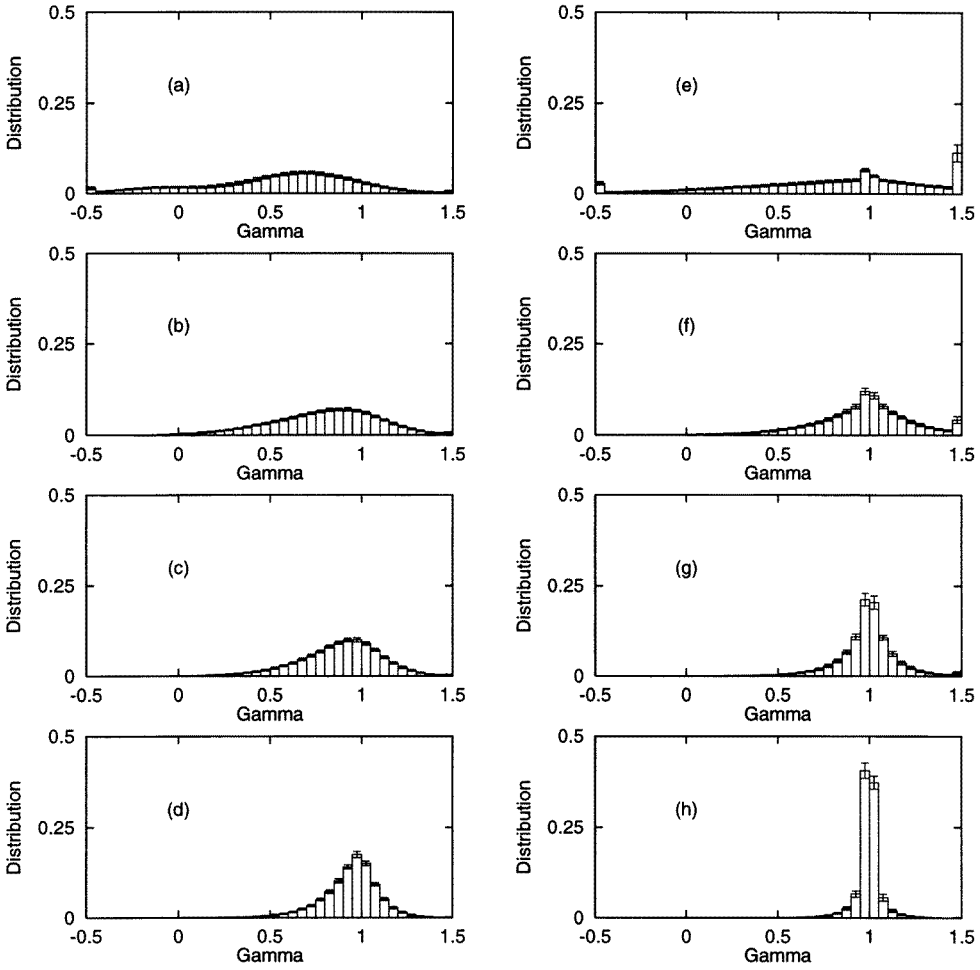

Figure 3. The average number of stability coefficients n_γ per interval of size 0.05, divided by the total number of the stability coefficients $\gamma_i^{\mu,m}$, given by NpL, has been plotted for a neural network with dilution 0.6, after one or more learning cycles, for the non-local and local energy saving learning rules. The calculations have been performed for a tabula rasa network, $w_{ij}(t_0) = 0$, of $N = 128$ neurons with vanishing thresholds ($\theta_i = 0$). An average has been taken of $L = 100$ sets of $p = 32$ patterns. The average activity is $a = 0.2$. (*a*)–(*d*) The average number of stability coefficients after 1, 5, 10 and 20 learning cycles in case of the local energy saving learning rule (66), (67). (*e*)–(*h*). The average number of stability coefficients after 1, 5, 10 and 20 learning cycles in case of the non-local energy saving learning rule (42).

approximative, is an excellent competitor of the non-local one. After five learning cycles the number of negative $\gamma_i^{\mu,m}$ is already negligible (see figures 3(*b*) and (*f*)), and the distribution of the $\gamma_i^{\mu,m}$ are comparable.

We finally make some observations regarding other learning rules. In view of (24), the symmetric learning rule (23) yields the same values of the γ as in the case of our asymmetric learning rule (20). Hence, in particular, the whole analysis of this section holds true for the symmetric learning rule as well. In other words, although the changes Δw_{ij} in the weights w_{ij} as given by the symmetric learning rule (23) are, of course, different from those given by our asymmetric learning rule (20), the convergence properties—studied here via the γ—are

exactly the same for the symmetric learning rule (23) and our asymmetric learning rule (20). The 'wrong' asymmetric learning rule (22) does not work at all, as has been explained at the end of section 3.

8. Summary

We have shown that two different arguments, a biological one (section 3 and a physical one (section 4) lead to a Hebb rule of the same asymmetric form: compare equations (20), (21) and (42). A learning rule of this form is never, or at least not often, used in the physical literature, which, in general, is less concerned with an accurate modelling of a biological network.

The biological argument was largely based on the improbability of a change of connections if the pre-synaptic neuron was inactive. The physical argument was based on the expression (35) for the energy change, not on any ad hoc cost-function like (27) as has been done so far in the literature. The local version of the energy saving Hebb rule (42), given by equations (66), (67), may be relevant for biological systems. It has been tested numerically in section 7, and turns out to yield storage of patterns in a satisfactory way (see in particular figure 3).

Acknowledgments

The authors are indebted to Bob van Dijk, Hugo Keizer, Hans Capel, Bernard Nienhuis, Wytse Wadman, Henk Spekreijse and the referees, all of whom contributed in some way to this paper in its present form.

Appendix A. Maximal efficient learning

We shall here merely verify the maximal efficient learning rule, not derive the rule, since the derivation closely parallels the one of Linkevich [8]. In view of the special constraints mentioned directly above (equation (63)), equation (12) reduces to

$$\sum_{j\in V_i} w_{ij}(t)\xi_j^\mu = 2\kappa\xi_i^\mu \qquad (\mu = 1,\ldots,p). \tag{A.1}$$

Similarly, the solution (57) of (12) reduces to

$$w_{ij}(t) = \begin{cases} N^{-1}\displaystyle\sum_{\mu,\nu=1}^{p} 2\kappa\xi_i^\mu (C_i^{-1})^{\mu\nu}\xi_j^\nu & (j\in V_i) \\ 0 & (j\in V_i^c). \end{cases} \tag{A.2}$$

In order to store a new pattern $\boldsymbol{\xi}^{p+1}$, the new weights $w_{ij}(t')$ have to obey the equations

$$\sum_{j\in V_i} w_{ij}(t')\xi_j^\mu = 2\kappa\xi_i^\mu \qquad (\mu = 1,\ldots,p+1). \tag{A.3}$$

The weights $w_{ij}(t')$ are related to the weights $w_{ij}(t)$ by

$$w_{ij}(t') = \begin{cases} w_{ij}(t) + \Delta w_{ij}(t) & (j\in V_i) \\ w_{ij}(t) & (j\in V_i^c) \end{cases} \tag{A.4}$$

where the $w_{ij}(t)$ are the connections after storage of the patterns $\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^p$ as given by equation (A.2) and the $\Delta w_{ij}(t)$ are given by (63).

Inserting (A.4) with (A.2) and (63) into the left-hand side of (A.3) yields

$$\sum_{j\in V_i} w_{ij}(t')\xi_j^\mu = \begin{cases} 2\kappa\xi_i^\mu + \sum_{j\in V_i} \Delta w_{ij}(t)\xi_j^\mu & (\mu = 1,\ldots,p) \\ 2\kappa\xi_i^\mu & (\mu = p+1). \end{cases} \tag{A.5}$$

The right-hand side of these equations is equal to that of (A.3) if

$$\sum_{j\in V_i} \Delta w_{ij}(t)\xi_j^\mu = 0 \qquad (\mu = 1,\ldots,p). \tag{A.6}$$

In order to show that (A.6) holds, we first decompose ξ_j^{p+1} according to

$$\xi_j^{p+1} = \begin{cases} \sum_{\mu=1}^{p} a^\mu \xi_j^\mu + \psi_j^{p+1} & (j \in V_i) \\ \sum_{\mu=1}^{p} a_j^\mu \xi_j^\mu + \psi_j^{p+1} & (j \in V_i^c) \end{cases} \tag{A.7}$$

where a^μ, a_j^μ and ψ_j^{p+1} have been taken such that†

$$\sum_{j\in V_i} \xi_j^\mu \psi_j^{p+1} = 0 \qquad (\mu = 1,\ldots,p). \tag{A.8}$$

Using (A.2) and (A.8) one may prove the auxiliary relation

$$\sum_{k\in V_j} w_{jk}(t)\psi_k^{p+1} = 0. \tag{A.9}$$

The proof of (A.6) is now straightforward. First, substitution of (63) into (A.6) yields

$$\sum_{j\in V_i} \Delta w_{ij}(t)\xi_j^\mu \propto \sum_{j\in V_i} \left[2\kappa\xi_j^{p+1} - \sum_{k\in V_j} w_{jk}(t)\xi_k^{p+1}\right]\xi_j^\mu \qquad (\mu = 1,\ldots,p). \tag{A.10}$$

Then, substituting the decomposition (A.7) in (A.10), and using (A.1), (A.8) and (A.9) we see that this expression vanishes, which proves (A.6). Hence, the left-hand side of (A.3) equals the right-hand side of (A.3) for a learning rule given by (63).

Appendix B. Modified method of the pseudo-inverse

Consider the p sets of N linear equations

$$\sum_{j=1}^{N} w_{ij}x_j^\mu = a_i^\mu \qquad (i = 1,\ldots,N;\ \mu = 1,\ldots,p) \tag{B.1}$$

where x_j^μ and a_i^μ are known constants ($j = 1,\ldots,N$; $\mu = 1,\ldots,p$). The N^2 unknowns w_{ij} are not determined as long as $p < N$. Let V_i be the subset of indices j with the property that w_{ij} is a solution of the set of equations (B.1), and let the complement of the set V_i with respect to the total set of indices $(1,\ldots,N)$, denoted by V_i^c, contain the indices j with the property that the w_{ij} have the pre-described constant values b_{ij}, i.e.

$$w_{ij} = b_{ij} \qquad (j \in V_i^c). \tag{B.2}$$

† In the case where all connections may change in time, the index sets V_i are all equal to the set of all indices. Then the equations (A.7) with $j \in V_i^c$ disappear and (A.8) amounts to the condition that the vector ψ^{p+1} is orthogonal to the vectors ξ^μ ($\mu = 1,\ldots,p$). Hence, in this particular case there are $p+N$ restrictions (A.7) and (A.8) for $p+N$ variables a^μ and ψ_j.

chosen in such a way that the system of equations (B.1) does not become incompatible. If the set V_i^c is empty, a solution of (B.1) can be obtained via the Moore–Penrose pseudo-inverse matrix [5, 6]. We want to obtain a solution for w_{ij} of (B.1), (B.2) in the case where V_i^c is not empty, and the pseudo-inverse matrix cannot be used directly. To that end, we construct a new set of equations, closely related to (B.1), (B.2), which can be solved via the pseudo-inverse. We refer to this construction as the *modified method of the pseudo-inverse*.

We first define a new set of variables $\tilde{w}_{ij}$ according to

$$\tilde{w}_{ij} = w_{ij} - b_{ij} \qquad (i, j = 1, 2, \ldots, N) \tag{B.3}$$

where b_{ij} are arbitrary in case $j \in V_i$. We then have

$$\tilde{w}_{ij} = \begin{cases} w_{ij} - b_{ij} & (j \in V_i) \\ 0 & (j \in V_i^c). \end{cases} \tag{B.4}$$

The under-determined set of pN linear equations (B.1), (B.2) can now be rewritten

$$\sum_{j \in V_i} \tilde{w}_{ij} x_j^\mu = \tilde{a}_i^\mu \qquad (\mu = 1, \ldots, p) \tag{B.5}$$

where

$$\tilde{a}_i^\mu = a_i^\mu - \sum_{j=1}^{N} b_{ij} x_j^\mu . \tag{B.6}$$

Note that (B.5) cannot be solved with the help of the pseudo-inverse, since the summation is only with respect to a restricted set of indices $j \in V_i$. We therefore consider a new set of pN linear equations, namely

$$\sum_{j=1}^{N} v_{ij} y_j^\mu = \tilde{a}_i^\mu \qquad (\mu = 1, \ldots, p). \tag{B.7}$$

The relation of (B.7) to (B.5) can be made clear by taking

$$y_j^\mu = \begin{cases} x_j^\mu & (j \in V_i) \\ 0 & (j \in V_i^c) \end{cases} \tag{B.8}$$

since then the set of equations (B.7) for the N^2 unknowns v_{ij} $(i, j = 1, 2, \ldots, N)$ becomes identical to the set of equations (B.5) for the unknown $\tilde{w}_{ij}$ $(i = 1, \ldots, N; j \in V_i)$. The equation (B.7) can be solved with the help of the pseudo-inverse. The solution reads

$$v_{ij} = \sum_{\mu,\nu=1}^{p} \tilde{a}_i^\mu (C^{-1})^{\mu\nu} y_j^\nu \qquad (i, j = 1, 2. \ldots, N) \tag{B.9}$$

where $C^{\mu\nu}$ is the usual correlation matrix [24]

$$C^{\mu\nu} = \sum_{k=1}^{N} y_k^\mu y_k^\nu . \tag{B.10}$$

If we use (B.8), the matrix $C^{\mu\nu}$ becomes what we have called the 'reduced correlation matrix', given by

$$C_i^{\mu\nu} = \sum_{k \in V_i} x_k^\mu x_k^\nu . \tag{B.11}$$

The modified correlation matrix takes into account the modifications in the usual correlation matrix due to the particular network architecture as dictated by the index set V_i. The solutions v_{ij} become, using (B.8),

$$v_{ij} = \begin{cases} \displaystyle\sum_{\mu,\nu=1}^{p} \tilde{a}_i^\mu (C_i^{-1})^{\mu\nu} x_j^\nu & (j \in V_i) \\ 0 & (j \in V_i^c). \end{cases} \tag{B.12}$$

Hence, the solution (B.12) turns out to be compatible with (B.4) for $j \in V_i^c$. Putting now

$$\tilde{w}_{ij} = v_{ij} \qquad (i, j = 1, 2, \ldots, N) \tag{B.13}$$

we have obtained a solution for (B.5), as follows by comparing (B.7) and (B.5). In this way we find, transforming back from $\tilde{w}_{ij}$ to w_{ij} with the help of (B.3), and substituting (B.6), the final result for the solution of the under-determined set of equations (B.1), (B.2):

$$w_{ij} = \begin{cases} b_{ij} + \sum_{\mu,\nu=1}^{p} [a_i^\mu - \sum_{j=1}^{N} b_{ij} x_j^\mu](C_i^{-1})^{\mu\nu} x_j^\nu & (j \in V_i) \\ b_{ij} & (j \in V_i^c). \end{cases} \tag{B.14}$$

We recall that the b_{ij} are arbitrary for $j \in V_i$, and prescribed for $j \in V_i^c$. Notice that the solution (B.14) is not unique because of the arbitrary constants b_{ij} ($j \in V_i$).

We want to solve (12) for a network with changing connections w_{ij} if $j \in V_i$ and non-changing connections if $j \in V_i^c$. Applying (B.14) with

$$\begin{aligned} x_i^\mu &= \xi_i^\mu \\ b_{ij} &= w_{ij}(t_0) \qquad (i, j = 1, 2, \ldots, N) \\ a_i^\mu &= \kappa(2\xi_i^\mu - 1) + \theta_i \end{aligned} \tag{B.15}$$

we obtain at once (57). We thus arrive at the observation that the energy saving solution (57) coincides with the solution (B.14), obtained with the help of the modified method of the pseudo-inverse.

References

[1] Buonomano D V and Merzenich M M 1998 *Annu. Rev. Neurosci.* **21** 149
[2] Marder E 1998 *Annu. Rev. Neurosci.* **21** 25
[3] Brown T H, Kairiss E W and Keenan C L 1990 *Annu. Rev. Neurosci.* **13** 475
[4] Hebb D O 1949 *The Organization of Behavior* (New York: Wiley)
[5] Personnaz L, Guyon I and Dreyfus G 1985 *J. Physique Lett.* **46** L359
[6] Kohonen T 1984 *Self Organization and Associative Memory* (New York: Springer)
[7] Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949
[8] Linkevich A D 1992 *J. Phys. A: Math. Gen.* **25** 4139
[9] Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
[10] Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
[11] Gerstner W, Ritz R and van Hemmen J L 1993 *Biol. Cybern.* **68** 363
[12] Müller B, Reinhardt J and Strickland M T 1995 *Neural Networks: An Introduction* (Berlin: Springer)
[13] Domany E, van Hemmen J L and Schulten K (ed) 1991 *Models of Neural Networks* (Berlin: Springer)
Domany E, van Hemmen J L and Schulten K (ed) 1994 *Models of Neural Networks II: Temporal Aspects of Coding and Information Processing in Biological Systems* (Berlin: Springer)
Domany E, van Hemmen J L and Schulten K (ed) 1995 *Models of Neural Networks III: Association, Generalization and Representation* (Berlin: Springer)
[14] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison-Wesley) p 23
[15] Kandel E R, Schwartz J H and Jessell T M 1991 *Principles of Neural Science* (London: Prentice-Hall) p 166
[16] Abelles M 1982 *Studies of Brain Function* (New York: Springer)
[17] Kinzel W and Opper M 1991 Dynamics of learning *Models of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer)
[18] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[19] Gardner E, Mertens S, Zippelius A 1989 *J. Phys. A: Math. Gen.* **22** 2009
[20] Amit D J, Gutfreund H and Sompolinsky H 1987 *Phys. Rev.* A **35** 2293
[21] Hopfield J J 1982 *Proc. Natl Acad. Sci., USA* **79** 2554
[22] Carnahan B, Luther H A and Wilkes J O 1969 *Applied Numerical Methods* (New York: Wiley) p 299
[23] Heerema M 1999 *PhD Thesis* Amsterdam, to be published
[24] van Hemmen J L and Kuhn R 1991 Collective phenomena in neural networks *Models of Neural Networks* ed E Domany E, J L van Hemmen and K Schulten (Berlin: Springer)