Learning internal representations in an attractor neural network with analogue neurons

Daniel J. Amit[1] and Nicolas Brunel
INFN, Sezione di Roma, Istituto di Fisica
Universita di Roma, La Sapienza, Ple Aldo Moro, Roma

Abstract

A learning attractor neural network (LANN) with a double dynamics of neural activities and synaptic efficacies, operating on two different time scales is studied by simulations in preparation for an electronic implementation. The present network includes several quasi-realistic features: neurons are represented by their afferent currents and output spike rates; excitatory and inhibitory neurons are separated; attractor spike rates as well as coding levels in arriving stimuli are low; learning takes place only between excitatory units. Synaptic dynamics is an unsupervised, analog Hebbian process, but long term memory in the absence of neural activity is maintained by a refresh mechanism which on long time scales discretizes the synaptic values, converting learning into an asynchronous stochastic process induced by the stimuli on the synaptic efficacies.

This network is intended to learn a set of attractors from the statistics of freely arriving stimuli, which are represented by external synaptic inputs injected into the excitatory neurons. In the simulations different types of sequences of many thousands of stimuli are presented to the network that performs indiscriminantly retrieval and learning. Stimulus sequences differ in preassigned global statistics (including time dependent statistics); in orders of presentation of individual stimuli within a given statistics; in lengths of time intervals for each presentation and in the intervals separating one stimulus from another.

We find that the network effectively learns a set of attractors representing the statistics of the stimuli, and is able to modify its attractors when the input statistics change. Moreover, as the global input statistics changes the network can also forget attractors related to stimulus classes no longer presented. Forgetting takes place only due the arrival of new stimuli. The performance of the network and the statistics of the attractors are studied as a function of the input statistics. Most of the large scale characteristics of the learning dynamics can be captured theoretically.

This model modifies a previous implementation of a LANN composed of discrete neurons, in features of networks of more realistic neurons. The different elements have been designed to facilitate their implementation in silicon.

[1]On leave of absence from Racah Institute of Physics

1 Introduction, motivations

A Learning Attractor Neural Network (LANN) composed of essentially discrete ± 1 neurons accompanied by organic, unsupervised learning dynamics on synaptic efficacies has been implemented in electronics[1, 2]. The motivation for the material implementation of such a network was to obtain a physical system that organizes itself to represent internally the global statistics of a natural (free) inflow of external input (stimuli). The learning process, which is unsupervised, hebbian dynamics on the synaptic efficacies, is expected to translate the large scale statistical features of the input stream into dynamical attractors of the neural dynamics of the network. These attractors constitute internal representations of the statistics. They are expected to be both robust and plastic, in that small, ever present fluctuations in the input should modify them slowly if at all, but they should follow the temporal changes of the large scale statistics in the input stream. Such large scale changes in the input stream may provoke the elimination of some attractors and the generation of new ones, expressing forgetting in a palimpsestic context.

The double dynamics of the target network is unsupervised not only due to the absence of a computational paradigm (see e.g. discussion in ref. [2]), but also in that there is no external control separating learning from retrieval, or computational phases. Learning dynamics is affected exclusively and continually by neural activities (rates). These, in turn, may vary depending on whether a stimulus is or is not afferent on the network, and which stimulus was last presented. The implemented network 'forgets' only due to the creation of new representations. Therefore, in the absence of arriving stimuli, synapses are stable indefinitely. This stability is achieved at the price of a refresh mechanism, which on longer time scales maintains a discrete set of stable values for a synapse, even though on the short time scale synaptic dynamics is analog.

Due to the refresh mechanism, learning related to long term memory becomes a random walk among the discrete states. The synaptic matrix formed by such a process results necessarily[3, 4] in a palimpsest memory with an extremely low memory span. In the context of the first implementation, only partial remedies to this problem are available. In the present implementation the network can, in principle, reach absolute maximal memory spans.

The first LANN has also pointed out several limitations of the discrete neural implementation. The new network is a step more realistic in the sense that a set of biological requisites are implemented in it. It turns out that most of the implemented biological features are quite desirable also from the functional, as well as the electronic points of view.

In the present LANN there are two types of neurons, excitatory and inhibitory; neurons are analog elements characterized by a current-to-rate gain function (allowing operation at low spike rates); a very low fraction of the excitatory neurons is activated by a given stimulus (low coding level); excitatory neurons are coupled by a plastic synaptic matrix, which is to be structured by learning; inhibitory neurons are coupled in a fixed way to and from the excitatory neurons, to maintain a low level of overall activity in the excitatory network. For simplicity, we use one a single inhibitory unit, representing the inhibitory network. It is merely a schematic way of demonstrating that following learning, attractor dynamics with low rates is manifested.

The corresponding advantages of the present network are: the separation into two types of neurons allows one to consider all synapses as positive and eliminates the need for four-quadrant multipliers[1]. It also reduces the sensitivity of the network's function to fluctuations in the fraction of active neurons per stimulus (see e.g. [27] and Section 6 below.) The analog nature of the neural variables allows the amplitude of the neural activities themselves to express in their rates the presence of a stimulus or its absence. It also allows the network to express unambiguously the fact that a given stimulus is unfamiliar, by leading the network to a state of no activity[5]. The option of low coding levels in the stimuli is the key to resurrecting optimal storage with palimpsest[4].

The operation of the network is similar to the first LANN: First an input string is generated by defining class representative patterns with a given coding level. Around each representative, a class is defined as a set of patterns with the same coding level but with a given range of variation about the representative. A stimulus sequence is determined by prescribing 1. the order of presentation of, randomly chosen, members of each class; 2. the distribution of time intervals for the presentation of each stimulus; 3. the distribution of time intervals between every two consecutive stimuli. A typical sequence includes about 200 presentations of stimuli from a given class. In this first study the class representatives are uncorrelated.

What we expect and indeed observe is that after a sufficiently long sequence of presentations the synaptic modifications stabilize reverberations (stable activity distributions) which correspond to the classes: if a member of a class is briefly presented, the network maintains an activity distribution corresponding to the class after the stimulus is removed. The speed of learning can be controlled by varying the transition probabilities between the synaptic states in the stochastic learning process (see e.g. ref. [2]). These persistent distributions (attractors) are the internal representations of the classes and hence of the statistics of the input string. They fluctuate somewhat as different members of a class cause somewhat different synaptic modifications and due to the fact that the learning process is stochastic. But as long as the classes remain fixed the attractors are rather robust, despite the interjection of many stimuli which belong to no class. These persistent, selective activity distributions are sustained by the learned synaptic matrix. If a given stimulus does not belong to any of the classes, it is not recognized by the network. The signature is that all activities decay to zero and the network remains silent[5, 6].

Since some synaptic modifications occur for the presentation of every stimulus, if the presentation of members of certain classes is interrupted, those classes are forgotten. Such classes are replaced by representations of new classes shown to the network.

In Section 2 we describe the network and its elements. In Sections 3 and 4 we concentrate separately on the neural dynamics with a fixed synaptic matrix, and

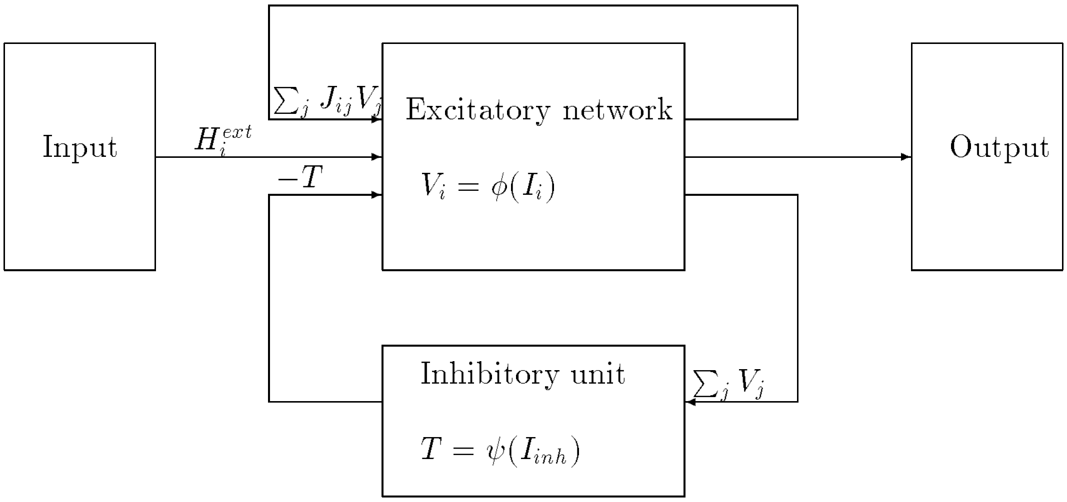

Figure 1: The LANN network. The excitatory network receives current from the input (stimulus), from the feedback collateral connections and from the inhibitory network. The computational outcome (attractor) is realized on the excitatory network.

on the synaptic stochastic dynamics with fixed neural variables, in order to develop the appropriate tools and intuitions. In Section 5 we present the simulations of the model with the double dynamics, i.e. when neurons are activated by a combination of external afferents and collaterals, and synaptic changes are driven by the resulting neural activities. In the last section we discuss the statistics of learned attractors.

2 The model

2.1 Composition of the network

The basic units of the network are N excitatory neurons i ($i = 1, \ldots, N$), characterized at time t by their incoming current $I_i(t)$ and their firing rate $V_i(t)$. A network composed of only excitatory elements is highly unstable, so we include inhibitory neurons to control the global activity in the excitatory network. If the synaptic weights to and from inhibitory neurons are fixed and uniform, the network of inhibitory elements can be represented by one equivalent inhibitory unit[12]. For simplicity, and in view of the electronic implementation, we use one inhibitory unit. Its activity (spike rate) is $T(t)$ and its synaptic input is proportional to the global activity in the excitatory network $A(t)$.

The connections between the different elements are

- Excitatory neurons are connected by plastic synapses $J_{ij}(t)$;

- The inhibitory neuron receives fixed uniform synapses from the excitatory network of strength K and sends its spike rate, $T(t)$, as inhibitory synaptic input to all excitatory neurons.

Fig. 1 presents schematically the different components of the network and their connection.

2.2 Dynamics of excitatory neurons

The dynamics of the afferent currents of excitatory neurons is described by the equation

$$\tau_{exc}\dot{I}_i(t) = -I_i(t) + \sum_{j\neq i} J_{ij}V_j(t) + H_i^{ext}(t) - T(t) \tag{1}$$

where τ_{exc} is the time constant of the currents charging the soma; $H_i^{ext}(t)$ is the external synaptic input (the stimulus); J_{ij} is the synaptic efficacy of the collateral synapse from neuron j to neuron i. It modulates the presynaptic afferent rate $V_j(t)$ to produce the synaptic input to neuron i; T is the synaptic input from the inhibitory neuron, which has the effect of a time-dependent threshold. A biological interpretation of these equations is given in refs. [8, 9, 10].

The firing rate of neuron i, $V_i(t)$, is determined by its mean afferent current (see e.g. [8]), via the transduction function ϕ_{exc}:

$$V_i(t) = \phi_{exc}\left(I_i(t)\right) \tag{2}$$

A realistic ϕ_{exc} would have the following properties: when the current is lower than the threshold θ_{exc} the firing rate, $V_i(t)$, tends rapidly to zero. Above threshold ϕ_{exc} is a continuous, increasing function, saturating at $V_i{=}V_{max}$, which would be of order of the inverse absolute refractory period, for very high currents. In the following all rates are expressed in units of V_{max}.

Since the inhibition prevents the rates of the neurons from approaching saturation, we choose below a function which does not saturate and is particularly convenient for the electronic implementation:

$$\phi_{exc}(I) = \begin{cases} g_{exc}\ln\left(\frac{I}{\theta_{exc}}\right) & \text{if } I > \theta_{exc} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

This simplified neural function is represented in Fig. 2a together with the transduction function of a leaky integrate-and-fire neuron [8]. The comparison allows a calibration of the rates of the simplified function. The scale of the rates in the function Eq. 3, in the absence of a saturation, is arbitrary. Thus we use the more complete function, in the low frequency region where it approximates well the simplified function, to justify reading the values of the simplified function (Eq. 3) as a fraction of the maximal firing rate. The replacement of the more complete (integrate-and-fire) function by the simplified one in the functional region where they have qualitatively similar shape will produce the same behavior. In fact, one could stabilize attractors in a network with both excitatory and inhibitory threshold-linear neurons[11].

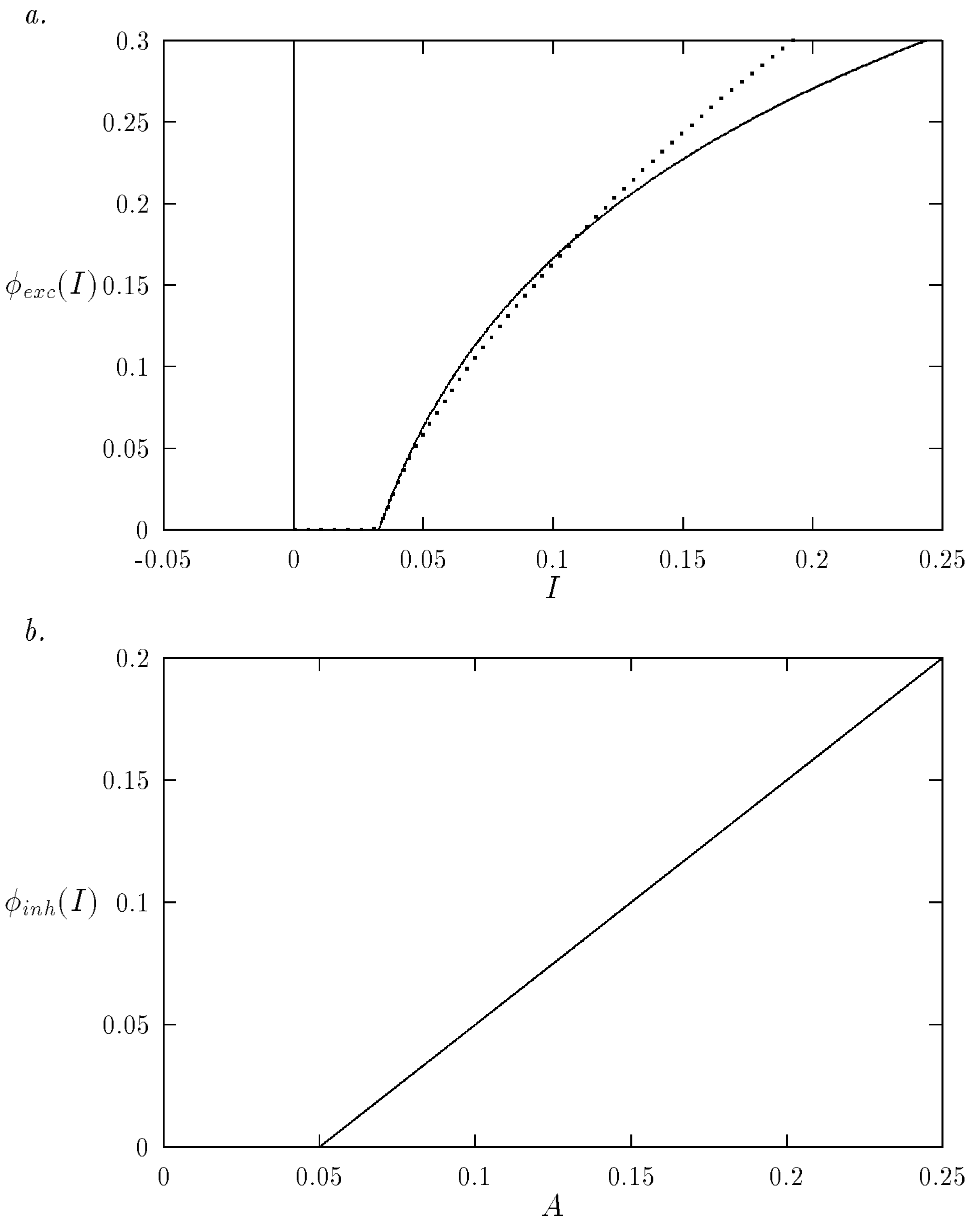

Figure 2: (a) Excitatory transfer function, with g_{exc}=0.15 and θ_{exc}=0.033 (full). Integrate-and-fire transduction function[8] for rate calibration (dotted), with parameters μ=0, σ=0.004 and θ=0.04. (b) Inhibitory transfer function: g_{inh}=1 and θ_{inh}=0.05.

Note that here it is not our objective here to reproduce rates as low as observed in cortical neurons.

2.3 Dynamics of the inhibitory element

The inhibitory neuron has a similar evolution equation

$$\tau_{inh}\dot{A}(t) = -A(t) + K\sum_j V_j(t) \tag{4}$$

where $A(t)$ is its afferent current and K is a uniform excitatory to inhibitory synaptic efficacy. The integration time constant τ_{inh} is chosen to be much shorter than the excitatory time constant, for the inhibition to effectively play its role. Since τ_{inh} will be the smallest time constant for the different elements, we will express all other time constants in terms of this unit. Thus we set τ_{inh}=1, and typically in the simulations τ_{exc}=5. The response of the inhibitory neuron $T(t)$ is

$$T(t) = \phi_{inh}(A(t)) \tag{5}$$

where ϕ_{inh} is the inhibitory transduction function. ϕ_{inh} is also a continuous increasing function that vanishes below the inhibitory threshold, θ_{inh}. For simplicity we take a threshold-linear function of the type

$$\phi_{inh}(A) = \begin{cases} g_{inh}(A-\theta_{inh}) & \text{if } A > \theta_{inh} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

with g_{inh}=1, θ_{inh}=0.05. It is shown in Fig. 2b. This simple transduction function models, in a schematic way, the reaction of a population of inhibitory neurons to an increase in its afferent excitatory currents. Its role is to represent qualitatively the inhibitory control of the activity in the excitatory network via the response T.

The inhibitory response T is then fed back as an inhibitory (hyperpolarizing) input to the excitatory neurons in Eq. 1. Since the source term in Eq. 4 is proportional to the global excitatory activity, when the global activity in the excitatory network exceeds the inhibitory threshold, the response of the inhibitory neuron increases linearly with the excitatory activity. In this way the inhibitory neuron provides an efficient control of the activity in the excitatory network and prevents the percolation of the excitation to the entire network[13]. Also the details of the inhibitory function are not essential. A different function with similar features produces essentially the same performance, except possibly for the absolute values of the stable rates.

2.4 Synaptic dynamics and long-term memory

A plausible form of the synaptic dynamics that is also naturally implementable in electronics, can be the following:

$$\tau_c\dot{J}_{ij}(t) = -J_{ij}(t) + c_{ij}(t) + J\Theta\left(J_{ij}(t) - w_{ij}(t)\right). \tag{7}$$

It is basically an integrator with a time constant τ_c. The integrator has a structured source $c_{ij}(t)$, representing hebbian learning. This source is given in terms of the neural rates, $V_i(t)$ and $V_j(t)$, of the two neurons connected by this synapse as

$$c_{ij}(t) = \lambda V_i(t) V_j(t) - \mu [V_i(t) + V_j(t)] \tag{8}$$

λ and μ are positive parameters separating potentiation from depression. Their values are chosen so that when the rates of both neurons are high $c_{ij} > 0$; if one is high and one is low $c_{ij} < 0$; and if both are zero, or very low c_{ij} is negligible.

The last term on the right hand side of Eq. 7 is the refresh mechanism discussed in detail in ref. [2]. It represents one way of preventing the loss of memory due to the decay of the integrator when no source is present. It functions in the following way: If at any given moment the source $c_{ij}(t)$, of any given synapse asynchronously, exceeds the fluctuating threshold $w_{ij}(t)$ a refresh source turns on to drive the synapse to the high value J. If the refresh charging rate is fast, on the time scale of the high frequency cutoff of the fluctuating threshold, the synaptic value will not turn back and that synapse will be driven to the value J. In the absence of a source this synaptic value will remain above its threshold and the efficacy J will be stable, indefinitely. On the other hand, if the instantaneous synaptic value is low, either because it started low, or because it was high and the learning source was negative enough, the refresh source turns off, asynchronously again, and in the absence of a source that synapse decays to zero. This is the other long-term, stable state of a synapse. The transition of a synapse from the lower stable state to the upper one is identified with LTP. The opposite transition is LTD.

As a consequence, each synapse has two asymptotically stable values, 0 and J, and the synaptic dynamics, which is analog on the short time scale, is converted to a stochastic process between discrete stable states. The conversion of the analog to the discrete dynamics and the stochastic process itself will be discussed in further detail in Section 4. In the simulations to be presented we restricted ourselves to the stochastic transitions between the stable synaptic states. The electronic implementation of the corresponding analog dynamics is discussed in ref. [2].

The corresponding stochastic process is defined as follows: at time t the activities of the neurons i and j are $V_i(t)$ and $V_j(t)$. They determine the value of $c_{ij}(t)$ given by Eq. 8. If $c_{ij}(t)$ is above a threshold for potentiation θ_+, and the synaptic efficacy J_{ij}=0, then $J_{ij} \rightarrow J$ with probability p_+. If $c_{ij}(t)$ below a threshold for depression θ_-, and J_{ij}=J, then $J_{ij} \rightarrow 0$ with probability p_-. For a detailed discussion of this point see e.g. Section 4. Effectively, potentiation ($0 \rightarrow J$) occurs with probability p_+ when both post and presynaptic neurons are significantly active. Depression ($J \rightarrow 0$) occurs with probability p_- when only one of the two neurons is significantly active.

The idea of the fluctuating threshold for synaptic transitions is one way of introducing stochasticity into learning. It has been argued [3, 4] that if neurons with equivalent mean activities deterministically provoke a synaptic modification, the resulting network's performance is exceptionally poor. Stochasticity may find its origin in the fact that neural activities, even when equivalently driven, fluctuate. The synapse then feels a fluctuating source. A convenient schematization of this fact, is to take non-fluctuating neural activities and having an effective fluctuating threshold for the long-term synaptic transition.

2.5 Analog learning and short-term memory

Focusing on the behavior of the network with the discrete values of the synapses, learned in the above stochastic process, overlooks the combined neural-synaptic dynamics driven by the analog synaptic values. Analog synaptic dynamics generates intermediate values for the synaptic efficacies. With the removal of the stimulus, and in the absence of neural activity in the network, the synaptic efficacies will relax exponentially to the discrete values of long term memory, with a time constant τ_c. However, when neurons are active and the synapses remain modified for a couple of τ_cs following the removal of the stimulus, the behavior of the network may be modified. Especially since the number of synapses modified by the analog, deterministic dynamics is much higher than the number of discrete transitions provoked by the low-probablity stochastic process, which may compensate for the fact that the analog synaptic changes are of smaller magnitude than the discrete changes. The effects related to these transient analog values we call short term-memory. For example when a stimulus is presented to the network all synaptic efficacies connecting a pair of active neurons will increase. Only a fraction of them will cross the threshold for potentiation and be driven to J. The synapses which do not will have enhanced efficacies during presentation and after the stimulus is removed. Similar (but negative) changes will occur for synapses connecting an active neuron to an inactive one.

The short term memory is of course correlated, in a hebbian way, with the last stimulus presented. This raises the question whether the transient synaptic efficacies could not support a neural activity distribution similar to that of the stimulus, which in turn may produce analog short term synaptic modifications in a self-reproducing way. Such a possibility may generate a stationary state in which neurons and synapses support each other to maintain the activity of the last stimulus. Experiment[20] shows that this is not a typical case. A stimulus not shown during the training stage of the experiment does not produce a lasting neural activity distribution following its removal. This does not exclude the possibility that the effect is totally absent. In simulations we have run we find that in some conditions, before attractor behavior is manifested in the absence of the short-term effects, the analog synaptic changes can stabilize an attractor, which is an attractor of the double dynamics.

In the present study we do not discuss or study in detail the short term memory effects. First, because dealing with the discrete, long-term synaptic part is the generic case. Second, a detaied study of the full analog double dynamics is extremely time consuming in simulations. It might better be done on the electronically implemented network. This approach is permitted since the long-term synaptic changes are not affected by the dynamics provoked by the short-term analog memory, since during the inter-stimulus interval no transitions occur.

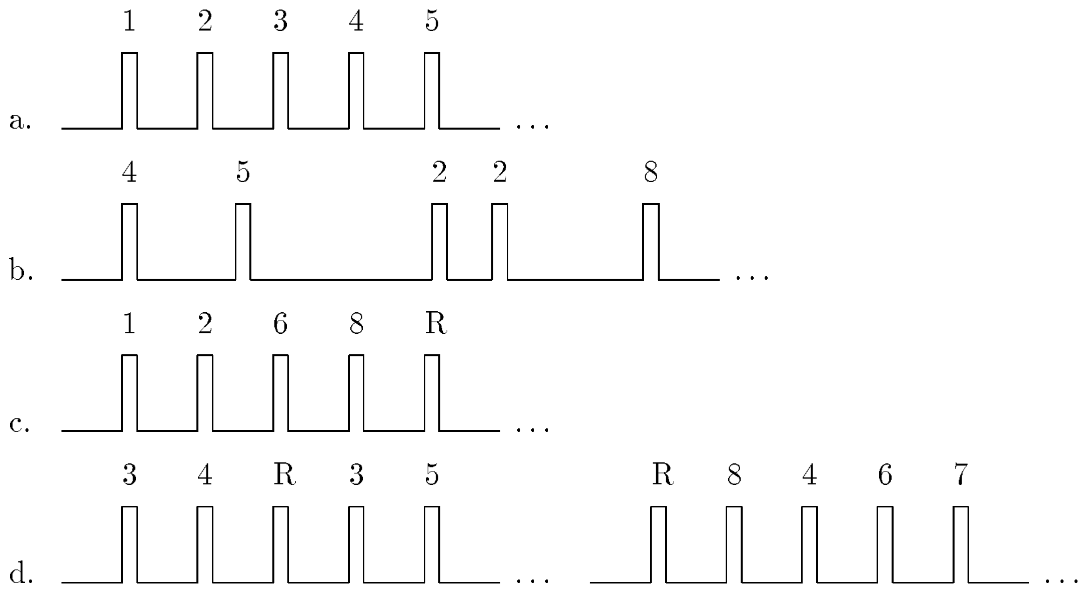

Figure 3: Stimulus stream: set of 9 prototypes. Each line shows the activation of the external synaptic inputs versus time during a particular run. The symbol above each bar indicates the class to which the stimulus belongs. In each example the presentation intervals are fixed. a. Fixed sequence, fixed inter-stimulus (IS) intervals. b. Random sequence, random IS intervals. c. Random sequence with random stimuli interjected (indicated as R), fixed IS intervals. d. Random sequence with varying statistics, fixed IS intervals: at the beginning only classes number 1-6 are shown. At the end only classes number 4 to 9 are shown.

2.6 External inputs — the stimuli

A stimulus is presented to the network as an injection of a distribution of synaptic inputs. Neuron i receives input $\{H_i^{ext}(t)\}$ (Eq. 1) lasting for an interval of length t_p, the presentation duration of the stimulus. Each presentation is followed by an inter stimulus interval of length t_r in which the external currents are set to zero. In a simulation a long sequence of stimuli (several thousands) is presented to the network separated by inter-stimulus intervals. The sequence of time interval lengths (stimulus and inter-stimulus) is either fixed or random. A stimulus is presented to the network at time t=0 as

$$H_i^{ext}(t) = \begin{cases} H\eta_i & \text{if } 0 < t < t_p \\ 0 & \text{if } t_p < t < t_p + t_r \end{cases} \tag{9}$$

where H is the strength of the external synaptic input, and $\{\eta_i\}$ is the N-bit word characterizing the stimulus.

2.7 Statistics of stimuli — classes

The statistics of the stream of incoming stimuli will determine how the network forms its synaptic matrix and thus its internal representations. In the simulations the stimuli are divided into two different types. Some stimuli belong to one of a set of predetermined classes, which are uncorrelated. Other stimuli are random and uncorrelated with any of the classes. The network is expected to create internal representations for each of the classes whose members are repeatedly presented. The other stimuli, i.e. those uncorrelated with any class, act as noise on the learning process.

At the beginning of the simulation each of p (about 30) classes are defined by a representative pattern, the prototype. Each prototype is an N-bit word $\{\xi_i^\mu = 0,1\}, i = 1,\ldots,N, \mu = 1,\ldots,p$, randomly and independently drawn from the distribution

$$P(\xi_i^\mu = 1) = f, \qquad P(\xi_i^\mu = 0) = 1 - f \tag{10}$$

The average number of activated neurons in each prototype is fN. f is the coding level of the prototypes, and will be typically low in the simulations — of the order 0.05–0.1. A neuron i for which ξ_i^μ=1 is said to be in the foreground of prototype μ. If ξ_i^μ=0, the neuron is in the background. Each prototype defines its corresponding class (or cluster) of stimuli. The members of a class are noisy versions of a prototype. To be specific, a pattern of class μ is a word $\{\eta_i\}$ chosen randomly in the following way:

- for foreground neurons, $\xi_i^\mu = 1$,

$$P(\eta_i = 1) = 1-(1-f)x, \qquad P(\eta_i = 0) = (1-f)x$$

- for background neurons, $\xi_i^\mu = 0$,

$$P(\eta_i = 1) = fx, \qquad P(\eta_i = 0) = 1 - fx$$

where x is a parameter measuring the average correlation between a pattern of a class and its prototype. This procedure ensures that the mean number of activated neurons per stimulus remains fN whatever the value of the parameter x. For $x = 0$ the stimulus is identical to the prototype, while if $x = 1$ it is statistically independent of it. x also measures the extent of a given class in stimulus space. If x is small, each class is a very small region around its prototype. When x increases the region defining each class grows. To each member of a class corresponds a stimulus $H_i^{ext} = H\eta_i$.

Next a sequence of presentations is defined (for examples see below). A sequence is composed of arbitrary members from each class, in an order defined for the classes. Some sequences include stimuli not belonging to any class. The network is expected to learn, in an unsupervised way, a representative attractor for each class. See e.g. [12, 14, 15].

Several presentation procedures have been considered. Some of them are represented schematically in Fig. 3:

- Fixed sequence (Fig. 3a): classes are always presented in the same order. This is the situation in the experiments of Miyashita[16].
- Random sequence (Fig. 3b): classes are selected at random at each presentation.
- Random sequence including random stimuli (Fig. 3c): stimuli are either members of a class selected at random or are uncorrelated with any prototype (class). They appear in the sequence with probability $(1-q)$ or q, respectively.
- Random sequence with time-varying statistics (Fig. 3d): the probability of presentation of a given class varies with time. Specifically, classes are divided in three groups, each containing the same number of classes $p_c = p/3$. During a first period $0 < t < T$ only stimuli from the first two groups of classes are shown to the network. T is long enough so that many members of each presented class appear. The probability of presentation of any class is $1/(2p_c)$. After this period the statistics change: for $t > T$ only stimuli of the second and third groups of classes are presented to the network, while stimuli of the first group are not shown anymore.

3 Neural dynamics with fixed synaptic matrix

The focus of attention here is on the effect of unsupervised learning on the formation of attractor representations in the network. In order to be able to evaluate the attractor performance in the learning network, we first describe the performance of the network in the more familiar situation in which the synaptic matrix is well formed and fixed. This will serve as a standard for good attractor dynamics.

3.1 Stable selective frequencies — analysis

We suppose that the matrix has already learnt the set of prototypes $\{\xi_i^\mu\}$, after they have been presented many times to the network. For simplicity we take the Willshaw matrix [18], i.e.

$$J_{ij} = \begin{cases} J & \text{if for at least one pattern } \mu \;\; \xi_i^\mu \xi_j^\mu = 1 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

and we choose the normalization $J = 1/(fN-1)$ ($fN-1$ is the number of excitatory connections from any foreground neuron to all other neurons in the foreground of a prototype with fN active neurons, because there are no self-connections. The -1 is kept since in our simulations fN is rather small.) The normalization of the excitatory to inhibitory coupling is chosen to be $K = 1/fN$. With these normalizations the total excitatory synaptic input to both excitatory and inhibitory neurons is independent of f and N when the network is in a configuration in which all foreground neurons of a given prototype are active, while background neurons are inactive. This allows testing systems with varying numbers of neurons and coding levels keeping all neural parameters such as the threshold fixed.

Such a synaptic matrix would be the result of the following simplified learning dynamics: 1. one starts with all synaptic efficacies at zero. 2. When a prototype is shown to the network a given synaptic efficacy J_{ij} is modified only if $\xi_i^\mu \xi_j^\mu = 1$. In this case the synaptic eficacy is potentiated and becomes equal to J. After all prototypes have been shown at least once the synaptic matrix will be given by Eq. (11). It would be a very idealized case of the synaptic dynamics of Eq. 7. First because it does not have the stochastic nature and second because it does not take into account the fact that the arriving stimuli may be noisy.

Suppose the network is in a configuration in which all neurons in the foreground of prototype number μ are active with frequency V, while all other neurons are quiet. The recurrent input in Eq. 1, coming from the other excitatory neurons, is

$$\sum_{j \neq i} J_{ij} V_j = \begin{cases} V & \text{if } \xi_i^\mu = 1 \text{ (foreground)} \\ V\Delta_i & \text{if } \xi_i^\mu = 0 \text{ (background)} \end{cases} \tag{12}$$

Δ_i has a distribution peaked around the mean value [18]

$$\Delta = 1 - (1 - f^2)^{p-1}.$$

If the number of stored prototypes is low, $pf^2 \ll 1$, we have $\Delta \ll 1$ and the distribution of foreground and background excitatory synaptic inputs will be separated by a large gap. If the threshold for the activation of the excitatory neurons is in this gap, it will enable the network to remain in a configuration where all foreground neurons are active while background neurons are silent — the attractor corresponding to the learned prototype. If the threshold is fixed the network will be very unstable — excitation can easily percolate to the whole network. The inhibitory neuron acts as an activity dependent threshold, which can be adjusted to be inside the gap, even when the location of this gap depends on the particular attractor.

If the network is a stable state with all foreground neurons active at frequency V, and all other neurons quiet, then Eqs. (1, 2, 4, 5) and (12) imply:

$$I_i = V - \phi_{inh}(V)$$

for the 'on', or foreground, neurons, and

$$I_i = V\Delta_i - \phi_{inh}(V) < \theta_{exc}$$

for the 'off', or background, neurons.

The stable selective frequency is thus the solution of the equation

$$V = \phi_{exc}(V - \phi_{inh}(V))$$

In Fig. 4 we draw the right and left hand sides of the above equation for four values of the inhibitory gain g_{inh}. The highest intersection of this function with a straight

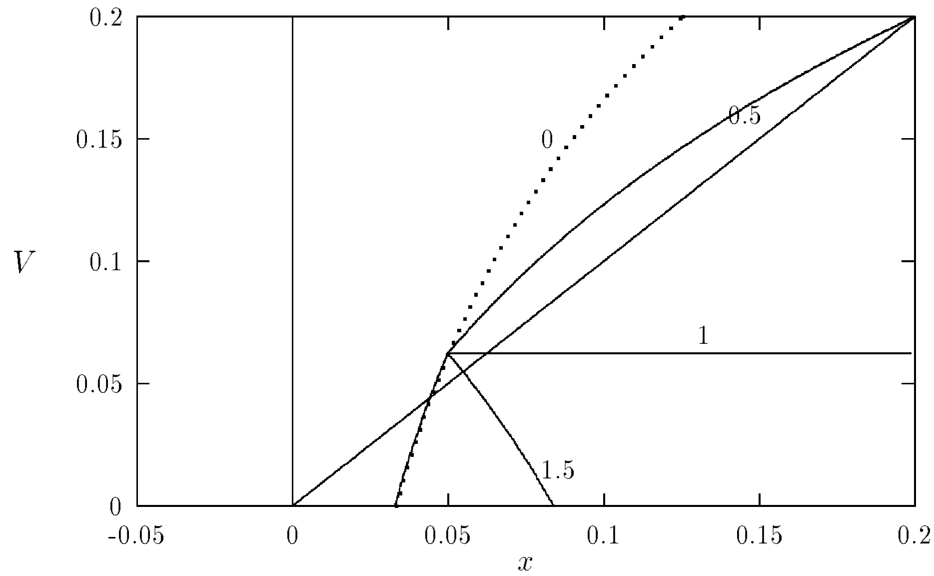

Figure 4: The function $x \to \phi_{exc}[x - \phi_{inh}(x)]$ with $\theta_{inh} = 0.05$ and four values of the inhibitory gain g_{inh}, indicated next to the corresponding curve. For each of these curves the highest intersection with a straight line of slope unity corresponds to the stable selective frequency. The parameters for the excitatory gain function are the same as in Fig. 2. The frequency is in units of saturation frequency.

line of slope unity corresponds to the stable frequency of the 'on' neurons in an attractor. It shows that inhibition can have a very significant effect on the frequency in the attractor. It decreases from 0.2 for g_{inh}=0.5 to around 0.05 for g_{inh}=1.5. Furthermore, it prevents the propagation of excitation to too many neurons, as we will see in the following section. In fact, the parameters of the transduction functions can be adjusted so as to have any desired value for the stable selective frequency.

Note however that Eq. 12 holds only if the number of active neurons in pattern μ is exactly fN. Otherwise the quantities on the right hand side have to be multiplied by f^μ/f, where f^μ is the actual fraction of active neurons in pattern μ, i.e.

$$f^\mu = \frac{1}{N}\sum_i \xi_i^\mu$$

If the prototypes are drawn randomly according to

$$P(\xi_i^\mu = 1) = f \quad P(\xi_i^\mu = 0) = 1 - f \tag{13}$$

f^μ is a fluctuating quantity with mean f and RMS

$$\sqrt{\frac{f(1-f)}{N}}$$

For N=200 and f=0.05 the RMS is about 0.015, which is about 30% of the mean. Thus the fluctuations of the number of active neurons per prototype will be relatively large[19]. In Secs. 4 and 5 we first use prototypes with a fixed number fN

of active neurons. In Sec. 6 we return to the question of these fluctuations and see how they affect the statistics of the learnt attractors.

3.2 Simulations

To study the attractor dynamics we have performed simulations of the network of N=200 neurons with fixed synapses, storing p=30 prototypes with a coding level f=0.05 (on average, ten active neurons per pattern). We used $\tau_{exc} = 5$. The condition $\tau_{exc} > 3$ was found to be necessary for the network to operate well. Stimuli were presented to the network (as in Eq. (9)) with H=0.1, unless otherwise stated, for a duration of t_p=100 (the time unit is the inhibitory time constant). Thus the presentation time is much longer than the scale of the neuronal dynamics. Then the stimulus was removed.

To monitor the neural attractor dynamics we measure the average activity of neurons which are indicated as active in the prototype of each class, i.e.

$$m_+^\mu(t) = \frac{1}{fN} \sum_i V_i(t)\xi_i^\mu \tag{14}$$

and the average activity of neurons which are in the background of the corresponding class

$$m_0^\mu(t) = \frac{1}{(1-f)N} \sum_i V_i(t)(1-\xi_i^\mu), \tag{15}$$

together with the mean total activity in the excitatory network and the activity of the inhibitory neuron.

Note that the overlap defined in Ref. [9] corresponds to the difference between these two parameters. If this difference is large, $m_+^\mu \gg m_0^\mu$, it means that the network state is very correlated with the prototype of the corresponding class. If $m_+^\mu \sim m_0^\mu$, and the overlap is small, the network state is uncorrelated with it.

Different situations have been studied. They are illustrated in Figs. 5, 6, 7 and 8.

Silent initial network, stimulus from a given class (Fig. 5): a noisy prototype is presented as a stimulus for $0 < t < t_p$. Before presentation of the stimulus all the neurons in the network are silent, $V_i = 0$ for all i. When the stimulus is presented, the mean activity in the foreground of the corresponding class increases to a steady state value of about 0.18. After removal of the stimulus at $t = t_p$, the network goes to the attractor corresponding to the prototype: all foreground neurons fire at an elevated frequency of about 0.07, while all background neurons are silent. The inhibitory reaction is proportional to the mean activity in the network and prevents neurons which are not in the foreground of the corresponding class to be activated.

Initial network in an attractor, stimulus from a different class (Fig. 6): the network is initially in an attractor corresponding to a given class, here class number 1. This is expressed by the elevated constant value of the corresponding m_+ (light curve for negative times). A stimulus from a different class is presented

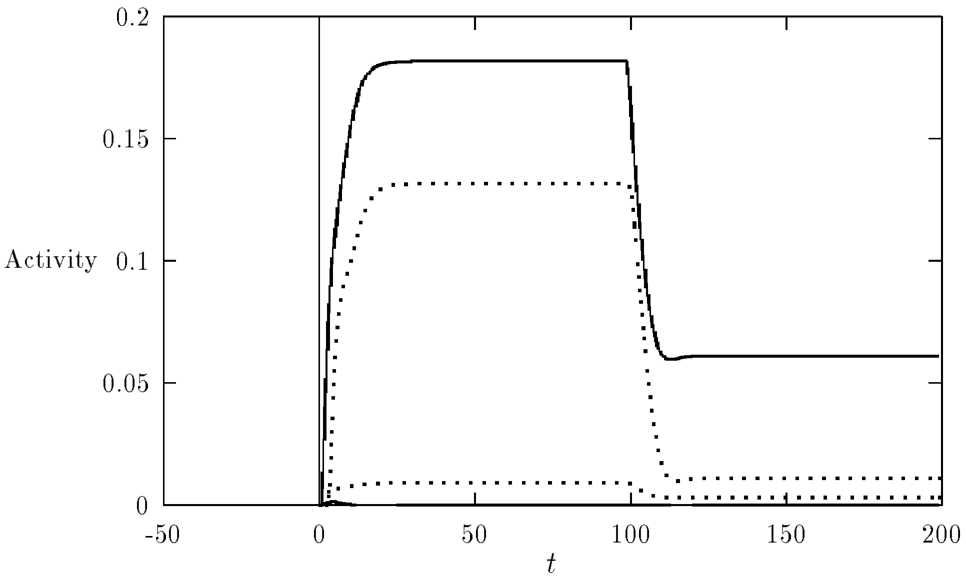

Figure 5: Attractor dynamics of initially silent network. Stimulus (noisy prototype) is presented between $t = 0$ and $t = 100$. Upper bold curve: mean activity of neurons belonging to the foreground of the prototype ($m^{\mu}_{+}(t)$). Lower bold curve (essentially along the axis): mean activity of the background neurons ($m^{\mu}_{0}(t)$). Upper dotted curve: activity of the inhibitory unit. Lower dotted curve: mean activity of all excitatory units.

at time 0: After a transient period in which neurons from both foregrounds are active, only the neurons tagged by the stimulus remain active, and the network evolves from the first attractor to the attractor corresponding to the second class. The transition takes place due to the inhibition, which prevents neurons in the foreground of the old prototype from staying active.

Initial network in an attractor, strong stimulus uncorrelated with any class (Fig. 7): a stimulus with high synaptic inputs, ($H = 0.1$), uncorrelated with any of the classes is presented for $0 < t < t_p$. The stimulus in this case destabilizes the previous attractor. After the stimulus is removed the network is unable to sustain any elevated activity.

Initial network in an attractor, weak stimulus (Fig. 8): a stimulus with low synaptic inputs, ($H = 0.01$) is presented for $0 < t < t_p$. The stimulus affects mildly the attractor and the network returns to the initial state after the stimulus is removed.

Note that the activity in the foreground depends on the number of active neurons of the particular pattern retreived, and is higher for a pattern with more neurons in the foreground. In all cases the activity in the background is zero and the overlaps with the other prototypes are low, of order f. The network is quite robust to noise in the presentation of the stimuli. It goes to the attractor corresponding to the prototype even if the noise level is high — until x is about 0.5.

When the double dynamics will be considered the synaptic matrix will of course

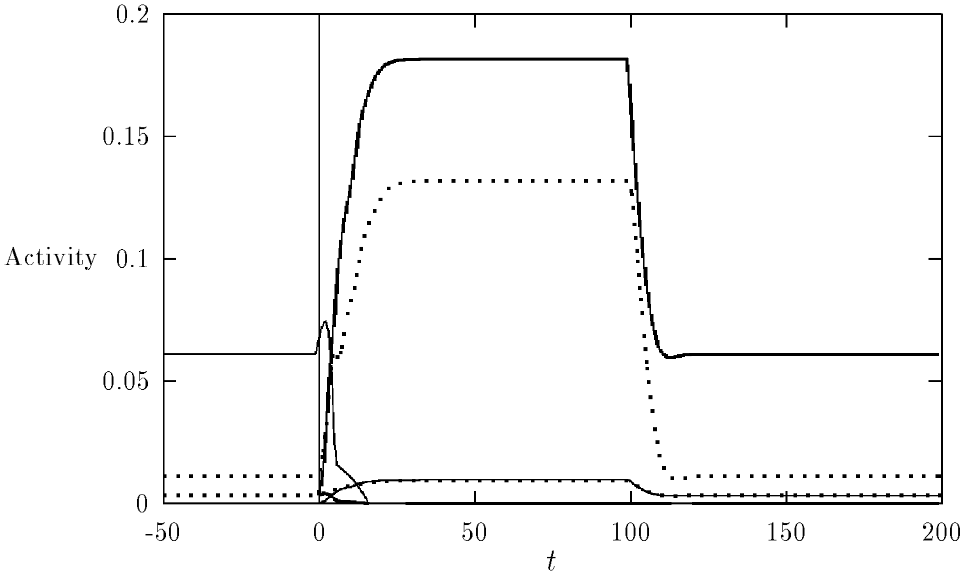

Figure 6: Network initially in an attractor corresponding to prototype 1 ($t < 0$). Stimulus from class 2 is presented for $0 < t < t_p$=100 with H=0.1. Light curves: mean activity of foreground and background of prototype 1. Bold curves: idem, prototype 2. Dotted curves: activity of inhibition (upper) and mean overall excitation (lower). Note that the mean activity in the excitatory network is the same for both attractors.

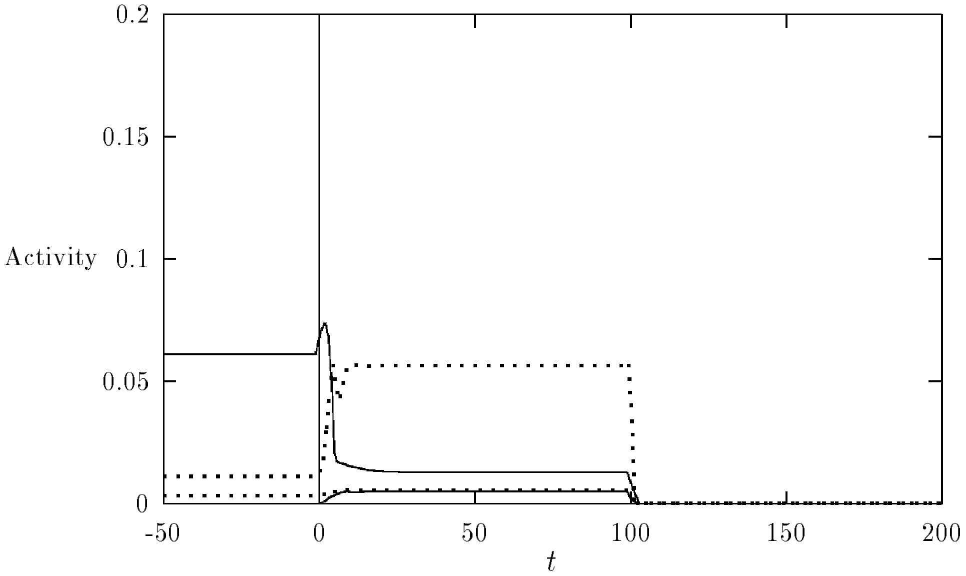

Figure 7: Network initially in an attractor ($t < 0$). Stimulus uncorrelated with any class presented for $0 < t < 100$ with H=0.1. All activity stops in the network. Light curves: mean activity in the foreground and background of prototype. Dotted curves: activity of inhibition and mean excitation.

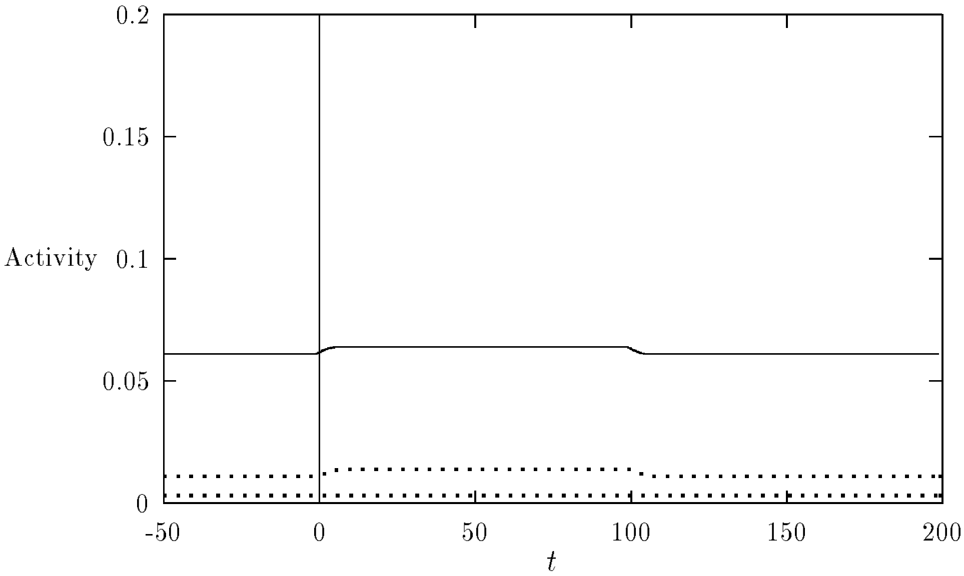

Figure 8: Network initially in attractor ($t < 0$). Weak stimulus uncorrelated with any class presented for $0 < t < 100$, with H=0.01. The network stays in the previous attractor. Light curves: mean activity in the foreground and background of prototype. Dotted curves: activity of inhibition and excitation.

be different from the matrix of Eq. (11). It will however retain some of its essential features, i.e. when a stimulus corresponding to a learned prototype is presented, the excitatory afferents coming to foreground neurons will be significantly larger than the ones coming to background neurons, and therefore there will be an attractor in which most of foreground neurons are active while most background ones stay silent. On the other hand the attractor, though very correlated with the prototype, will in general be different from it and fluctuate with time, due to noise in the presentations and the stochastic nature of learning.

4 Synaptic dynamics

4.1 Analog dynamics to transition probabilities

We turn now to isolated synaptic dynamics: neural rate variables are fixed. The rates can be high (externally imposed); elevated but low (as in an attractor in inter-stimulus intervals); or zero (for background). The double dynamics will be considered in the next section. First we define the stochastic learning process with a two-state synaptic efficacy which is to replace in the simulations the analog synaptic dynamics of Eq. (7).

We suppose that on long time scales neural variables can have one of three discrete activity states and the synapses two efficacy states. The rate variable V_i of neuron i can take on the following values:

- $V_i = V$ if the neuron is activated by a stimulus;
- $V_i = v$ if the neuron is active in an attractor in the absence of a stimulus;
- $V_i = 0$ in all other cases.

The ratio V/v, for fixed network parameters, depends on the amplitude of the imposed external input. For example, for the parameters of Section 3, we had $V/v \sim 5$ for $H = 0.2$, and 2.5 for $H = 0.1$ (see e.g. Fig. 5).

In the experiments of the Miyashita group in inferotemporal cortex [20, 16, 21] one can identify a qualitatively similar situation (see also [22]). Fluctuations around the three schematic values taken here are modelled by the stochastic behavior of the learning process.

For any given, fixed network state two neurons connected by a synapse can take only one out of seven different pairs of values: (V, V), (v, v), $(V, 0)$, $(0, V)$, $(v, 0)$, $(0, v)$ and (0,0). The pairs (V, v) and (v, V) can apear only during brief transients and are excluded from the discussion. The corresponding discretized values of the learning source term

$$c_{ij} = \lambda V_i V_j - \mu(V_i + V_j)$$

are:

- If both neurons are active:

$$c_{ij} = \begin{cases} C_+ = \lambda V^2 - 2\mu V & \text{during presentation of the stimulus} \\ c_+ = \lambda v^2 - 2\mu v & \text{in an attractor} \end{cases} \tag{16}$$

- If only one of the two neurons is active

$$c_{ij} = \begin{cases} C_- = -\mu V & \text{during presentation of the stimulus} \\ c_- = -\mu v & \text{in an attractor} \end{cases} \tag{17}$$

- If both neurons are silent

$$c_{ij} = 0.$$

The parameters λ, μ and H are chosen such that

$$C_+ < \langle w \rangle$$

and

$$1 + C_- > \langle w \rangle$$

where $\langle w \rangle$ is the average value of the fluctuating threshold. See e.g. [3, 4]. The RMS of w_{ij} will be denoted by $\langle \Delta w \rangle$. Transitions will then occur in two cases:

Potentiation: The initial synaptic efficacy is below threshold. If both neurons i and j are activated by the stimulus the analog synaptic value is driven by a learning source c_{ij}. It increases and eventually saturates at C_+. Here the synaptic value remains, as long as it is driven by the stimulus. In this interval it may be crossed by the fluctuating threshold, provided the saturation value is within the range of the fluctuations. The probability of a transition, p_+, will therefore depend on the ratio

$$\frac{\langle w \rangle - C_+}{\langle \Delta w \rangle}$$

the synaptic integration time τ_c, and the time of presentation of the stimulus t_p. These are LTP transitions.

Depression: The initial synaptic efficacy is above threshold and hence assisted by the refresh mecanism. If one of the neurons is activated by the stimulus and the other stays quiet, the analog synaptic value is driven by a negative learning source c_{ij}, superposed on the refresh source. The synaptic value decreases, and saturates at $1 + C_-$. It eventually crosses the threshold from above. The refresh turns off and a transition $J \to 0$ takes place with probability p_-, which depends on the ratio

$$\frac{1 + C_- - \langle w \rangle}{\langle \Delta w \rangle};$$

on the synaptic integration time τ_c, and on the time of presentation t_p. These are LTD transitions.

In all other cases long lasting transitions (LTP, LTD) will not take place.

In the present system the fluctuating threshold w_{ij} of the analog synaptic dynamics (Eq. (7)) is bounded away from 0 and J. As a consequence, if at least one of the neurons is inactive or both are active at low levels (v), the analog learning source c_{ij} is small or even negative, leading to a negligible probability that the refresh turn on if it had been off. If the synapse started from a low efficacy, following the removal of the stimulus, the synaptic efficacy returns to its original value. If the initial synaptic value is high, the refresh source is on. If both neurons are active; both are inactive; or one inactive and one active at low (attractor) level, the analog learning source is either positive or too weak to bring the synapse under threshold. The refresh mechanism remains on and following the removal of the stimulus, synapse regains its original value. In other words, the source term must go above a threshold θ_+ to allow an LTP transition and below a threshold θ_- for an LTD transition. There are therefore two such thresholds characterizing the source of the synaptic learning noise, as mentioned in Section 2.4.

Thus, in particular, the network can remain in an attractor, in the absence of a stimulus for an indefinite time without destroying past memory. See e.g. ref. [17]. This way we dispose of the artificial substraction terms introduced in ref. [1, 2]. See also e.g. ref. [17].

To summarize the qualitative properties of the transition probabilities in a concise way we fix the pre-synaptic activity at its high value and consider the probabilities as a function of the post-synaptic activity. Note that in this model pre and post synaptic neurons play a symmetric role in learning. For low post-synaptic activity there is a small depression probability for the strong synapses. In a wide range of intermediate values the probability for either potentiation or depression is essentially zero. As the post-synaptic activity increases the potentiation probability for weak synapses begins to increase monotonically with the post-synaptic activity. The qualitative behavior of the transition probabilities is shown in Fig. 16 of Ref. [2]. The zone where no transitions occur ensures the stability of long term memory. Interestingly, experiments in visual cortex suggest similar features for synaptic plasticity. See e.g. Ref. [23].

When the duration of presentation of the stimuli is kept fixed, learning can be approximated by a discrete stochastic process with two-state synapses. (See e.g. ref. [4]). The main lessons are:

- The performance of the network, in terms of either number of stored patterns or of quantity of information stored in the synaptic matrix, increases when the coding level f decreases.
- The capacity of the network is optimal, i.e. the number of patterns that the network can recall is of order $N^2/(\ln N)^2$, when $f \sim N/\ln N$ and

$$p_- \sim f p_+ \tag{18}$$

  i.e. when the numbers of up and down transitions balance each other. A similar relation is suggested in Ref. [24] to account for experimental findings.

In the simulations described below we have taken a low coding level, $f = 0.05$, and the transition probabilities satisfy Eq. 18. Note that the coding level cannot be decreased arbitrarily in the simulations since fN, the typical number of active neurons in an attractor, has to be large enough to ensure its robustness to noise.

4.2 Simulation of synaptic dynamics

We study the synaptic dynamics in the following way: a fixed sequence of p prototypes $\{\xi^\mu\}$ ($\mu = 1, \ldots, p$) with coding level f is generated. These define the classes. At each presentation the stimulus is a member of a class (in the sense defined in Section 2.7) selected at random. Upon presentation of a stimulus the synapse J_{ij} is modified according to the following rule:

- If J_{ij}=0 and both pre and post synaptic neurons are activated by the stimulus, V_i=V_j=V, then J_{ij} is set to 1 with probability p_+;
- If J_{ij}= 1, and only one of the pre and post synaptic neurons is active V_j=V, J_{ij} is set to zero with probability p_-;
- In all other cases the synaptic value remains unmodified.

This stochastic process approximates the dynamics described in Section 4.1 if

- the duration of presentation of each stimulus t_p is fixed,
- $t_p \gg \tau_c$ (the synaptic time constant) to allow the analog learning to reach saturation,
- the neural activities imposed by the external synaptic inputs remain constant throughout presentation,
- the time between presentations is long enough to allow the synapses to relax to their clipped values, $t_r \gg \tau_c$.

We simulate a network with the same parameters as in Section 3, i.e. N=200 neurons, f=0.05 (10 active neurons per prototype), p=30 classes. We choose transition probabilities p_+=0.1, p_-=0.005 and class size parameter x=0.1. Note that with the parameters selected, up and down transition rates are equilibrated, since $p_- = fp_+$. In a simulation a stream of 6000 stimuli is presented to the network, i.e. about 200 members of each class. In order to monitor the synaptic dynamics we observe the temporal evolution of the following observables:

- the average connectivity of the network $c(t)$

  $$c(t) = \frac{1}{N(N-1)} \sum_{i \neq j} J_{ij}(t),$$

  which is the fraction of potentiated synapses.

- The similarity of the learned synaptic matrix to the Willshaw matrix, for which we have tested the neural dynamics in Section 3. It corresponds to learning the same prototypes with parameters $p_+ = 1$, $p_- = 0$ and $x = 0$. This observable is defined as

  $$m_W(t) = \frac{1}{c_W N(N-1)} \sum_{i \neq j} J_{ij}(t) W_{ij}$$

  where W_{ij} is given in Eq. (11). c_W, the connectivity of the Willshaw matrix is given by,

  $$c_W = \frac{1}{N(N-1)} \sum_{i \neq j} W_{ij}.$$

  If the learned matrix is a random matrix, $m_W = c$, the connectivity. While if the learned matrix becomes identical to the Willshaw matrix, $m_W = 1$. The Willshaw matrix corresponding to the set of prototypes underlying the stimulus stream is considered as a standard with which learning is to be confronted.

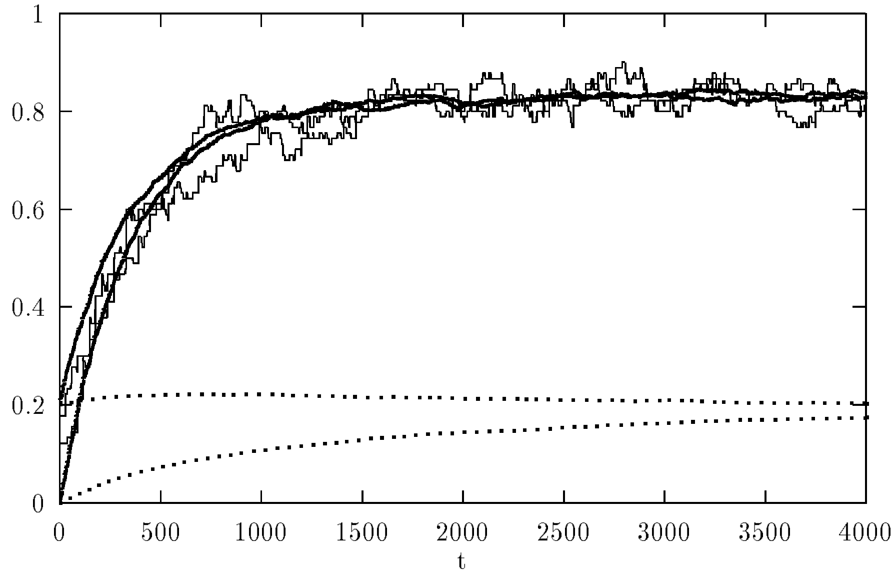

Figure 9: Evolution of the connectivity c (dotted curves), the overlap with the Willshaw matrix m_W (bold curves) and the intra-class connectivity c_μ showing how well a particular class (here class number 1) is learnt (light curves). $p = 30$ classes, $p_+ = 0.1$, $p_- = 0.005$, $x = 0.2$, and two initial values for the connectivity $c(0) = 0$ (lower curves at the origin) and $c(0) = 0.2$ (upper curves at the origin). Fixed sequence of classes.

- The average synaptic value for connections inside a population of neurons active in a given prototype:

$$c_\mu(t) = \frac{1}{fN(fN-1)} \sum_{i \neq j} J_{ij}(t) \xi_i^\mu \xi_j^\mu \tag{19}$$

  This observable measures how well a particular memory has been learned. For a random synaptic matrix (i.e. no learning at all) $c_\mu = c$, the connectivity. If all connections between the foreground neurons have been potentiated, $c_\mu = 1$. Note that for the Willshaw matrix we have $c_\mu = 1$ for all the classes. We will refer to this observable as intra-class connectivity (ICC).

The initial synaptic matrix is taken to be a random distribution of unit efficacies with connectivity (probability) $c(0)$. In Figs. 9 and 10 we show the intra-class connectivity of a particular class, here class number 1, the connectivity c, and the parameter m_W as a function of time. Each figure shows two runs corresponding to two different initial synaptic matrices: one with $c(0) = 0$, i.e. there are no synaptic links, the other with $c = 0.2$. In the latter case the initial synaptic values are chosen to be 1 independently at each synaptic site with probability 0.2.

We see that m_W and c_1 increase rapidly with time and after $t \sim 1000$ presentations, these parameters fluctuate around 0.8, regardless of the initial synaptic matrix. This corresponds to the presentation of about 30 members of each class.

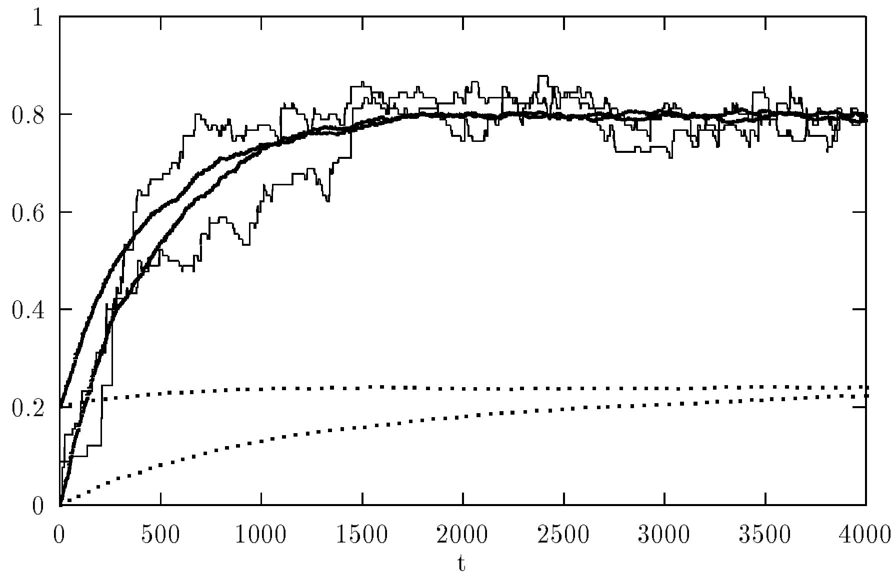

Figure 10: Same as Fig. 9, random sequence with probability $q = 0.2$ of stimuli uncorrelated with any class.

Fig. 9 corresponds to a fixed sequence of classes, while Fig. 10 corresponds to a random sequence of classes interspersed with a fraction $q = 0.2$ of random stimuli. In both cases the parameter measuring the spread of a class is $x = 0.2$. One can see that learning is a bit slower in the second situation, and fluctuations in the stationary state somewhat larger. This is caused by the presentations of the unclassified stimuli to the network. However the asymptotic mean values of the learning parameters are very close in both situations. This shows the robustness of the synaptic dynamics to presentations of unclassified stimuli.

The average intra-class connectivity (ICC) can be calculated analytically[25] in the case of multiple presentations of the same class, when random stimuli or members of other classes are presented between two presentations of this class. The calculation of the ICC with the parameters used in the simulations yields

$$c_1 \sim 0.8$$

i.e. a value very close to the asymptotic values of the ICC obtained in both Figs. 9 and 10.

5 Double dynamics

5.1 Dynamics with fixed input statistics

Finally we simulate the double dynamics of the network. The network is essentially the same as in Secs. 3 and 4, i.e. there are N=200 neurons, the neuronal time constant is τ_{exc}=5. The stimuli have a coding level f=0.05, and are generated in p=20 classes of extension x=0.1. Each stimulus shown is chosen to be either a random member of a randomly chosen class, or a configuration uncorrelated with any class. We chose q=0.1, i.e. 10% of the stimuli are uncorrelated with any prototype.

The duration of presentation of different stimuli is randomly chosen with a flat distribution. For each stimulus the presentation duration is chosen between 0 and 200, with an average at t_p=100. The duration between two presentations is also random. It is drawn from the same distribution as the presentation duration. Its average is t_r=100. Recall that all times are expressed in unit of the inhibitory time constant. Thus the average presentation duration as well as the average inter-stimulus interval are much longer than the neuronal time scales. The average presentation duration is the same as the chosen presentation duration in Sec. 3. Here we use random stimulus and inter-stimulus intervals in order to test the robustness of the network to another source of stochasticity. The choice of the distributions was motivated by its simplicity.

The mean interval between the beginnings of two presentations is

$$t_c = (t_p + t_r).$$

We choose the dependence of the transition probabilities on the presentation duration to be linear, i.e.

$$q_{\pm}(t) = \frac{t}{t_p} p_{\pm} \tag{20}$$

where $p_{\pm}$, the average transition probabilities, are taken to be the same as in the Sec. 4, i.e. p_+=0.1 and p_-=0.005. In the entire range of presentation durations both probabilities remain small, they can at most become double the above values. The above linear dependence expresses the idea that after the analog synaptic value has reached saturation corresponding to the current stimulus, it remains at saturation for the remaining duration of the stimulus. Note that in choosing the time dependence as in Eq. 20, we have neglected the time required for the analog term to reach saturation.

We have chosen the parameters in the learning source c_{ij} (Eq. 8) λ=1 and μ=0.025. Here LTP transitions occur only for $c_{ij} > \theta_+$, with θ_+=0.005. This implies that for equal pre and post synaptic activities, both activities must be greater than V=0.1 to induce transitions. LTD transitions occur for $c_{ij} < \theta_-$ with θ_-= -0.0025. If one neuron is silent, the other neuron's activity should be greater than V=0.1 to induce transitions. Note that in our simplified situation, with essentially bimodal activity distributions, there will be a single value to c_{ij} above the LTP threshold and a single value above the LTD threshold. This eliminates the need to consider the dependence of the transition probabilities on the value of c_{ij} (see e.g. ref [1]).

The simulation algorithm is as follows:

The initial conditions are:

- All neurons are at rest, $V_i(0) = 0$ for all i, and $T(0) = 0$;

- The synaptic matrix is initially random. Each synaptic link is present with probability 0.1, independently at each site. The initial connectivity is therefore $c(0) \sim 0.1$.

1. At the beginning of the presentation of every stimulus we draw:
   - The presentation duration;
   - The structure of the stimulus $\{\eta_i\}$, as described in Sec. 2.7.
2. The analog neural variables evolve according to Eqs. (1, 2, 4, 5), in which the external synaptic inputs are fixed to $H = 0.1$ if η_i=1 or to 0 otherwise, during the presentation interval. The basic integration step is Δt=0.1.
3. At the end of the presentation we calculate the value of c_{ij}, Eq. (8), at each synaptic site [2]. If at a given synaptic site c_{ij} exceeds the threshold θ_+, and $J_{ij} = 0$, a transition $0 \rightarrow J$ occurs with probability q_+ given by Eq. (20) where t is the presentation time. If c_{ij} is lower than θ_-, and $J_{ij} = J$, depression occurs with probability q_- given by Eq. (20).
4. At the beginning of the inter-stimulus interval we draw the length of the interval. During this interval neural variables evolve as in step (2), but with $H_i = 0$ for all neurons.
5. Go to step 1.

We monitor the learning and recall dynamics by measuring the synaptic observables (defined in Section 4): m_W – the similarity of the learned synaptic structure to the Willshaw matrix; c_μ – the intra-class connectivity for a particular class, μ=1 and the overall connectivity c. In the following we show the ICC for only one class, since all classes are presented with identical statistics and therefore the ICCs have a similar time course. The fluctuations from class to class are of the same order as those observed in Figs. 9 and 10. Moreover, as in Section 4.2, the intra-class connectivity fluctuates little about the observable m_W.

To monitor the neural attractor dynamics we choose a particular class. Every time a member of that class is presented, we wait until the end of the inter-stimulus interval following removal of that stimulus. At this point we measure the average activity of neurons which are indicated as active in the prototype of the corresponding class m_+^μ given in Eq. (14), and the average activity of neurons which are in the background of the corresponding class m_0^μ, Eq. (15)

One expects on the basis of the discussion in Section 3 that if the corresponding class has not been learned, all neurons in the network will become inactive and m_+^μ=m_0^μ=0. After a class is learned, the network should flow to an attractor

[2]Because a very small fraction of the synapses changes upon a single presentation, these changes have a negligible effect on the neural dynamics during the presentation, and the precise instant of the modification is immaterial. So we choose it at the end of each presentation for convenience.

correlated with the class prototype, and m_+^μ should attain a sizable value, while m_0^μ should remain negligible. Thus the quality of learning is expressed in both the synaptic observables c_μ, Eq. (19) and the neural observables $m_{+,0}^\mu$, Eqs. (14, 15). In fact, in all cases in which attractor behavior was observed, m_0^μ has been too small to be distinguished from the axis.

The mean time between the presentation of two successive stimuli is t_r defined above. We define a cycle as the period between two presentations of a stimulus of the same class. It has a mean duration of $p(1+q)t_c$. The duration of an entire simulation is typically $200p(1+q)t_c$, i.e. each class is presented about 200 times. Because of the presence of two different time scales we present the following figures with a time unit of t_c ($200\tau_{inh}$).

We show in Fig. 11a the evolution of the synaptic observables introduced in section 4. It is seen that the evolution of these parameters is rather similar to that observed in the case of the synaptic dynamics with fixed neural activities. In Fig. 11b we show the evolution of the selective delay activity m_+^1, i.e. the mean activity in the foreground neurons of a class number 1, after the delay following the presentation of a stimulus from this particular class, as explained above. This observable is a measure of the retrieval, or recall, quality. The mean activity in the background neurons, m_0^1, stays essentially at zero throughout the entire simulation. See e.g. Section 3.2 for discussion of these variables.

The two figures should be viewed in conjunction. Note first that learning changes only slightly the overall connectivity in the network. This is due to the choice of matched LTP and LTD (Eq. 18). Both the similarity to the Willshaw matrix and the intra-class connectivity (ICC), of class number 1, rise relatively rapidly and c_1 reaches the level of 0.7 after about $600t_c$, which amounts to about 30 cycles (a period in which about 30 members of each class are presented to the network). It is at this point that attractor behavior (delay activity) sets in a stable way. After this point, following the presentation of each stimulus, an activity distribution correlated with the prototype persists at a mean rate of 0.05–0.08 of saturation. An attractor correlated with the prototype has been formed in the synaptic structure and the class is recognized at every new presentation.

Several conclusions can be drawn from these figures: First, unfamiliar stimuli lead to the quiescent attractor. This is read from Fig. 11b in which one observes that during the first 30 presentations of stimuli of a given class the activity of foreground neurons is zero.[3] A second conclusion is that delay activity is not a single shot learning. One might have suspected that the synaptic structure that maintains the selective rate distribution accross the delay had been formed during the last presentation of the stimulus. This again is excluded by the fact that for the first 30 presentations of stimuli of a given class such attractor behavior is absent.

Just near the critical number of presentations (which depends on the particular set of parameters chosen) attractor behavior sometimes appears and later on the recall of the selective rate distribution follows every presentation of a stimulus. The

[3] In biology such neurons would be operating not at zero rates but at spontaneous activity levels of a couple of spikes per second. This can be captured also by network models. See e.g. ref. [6].

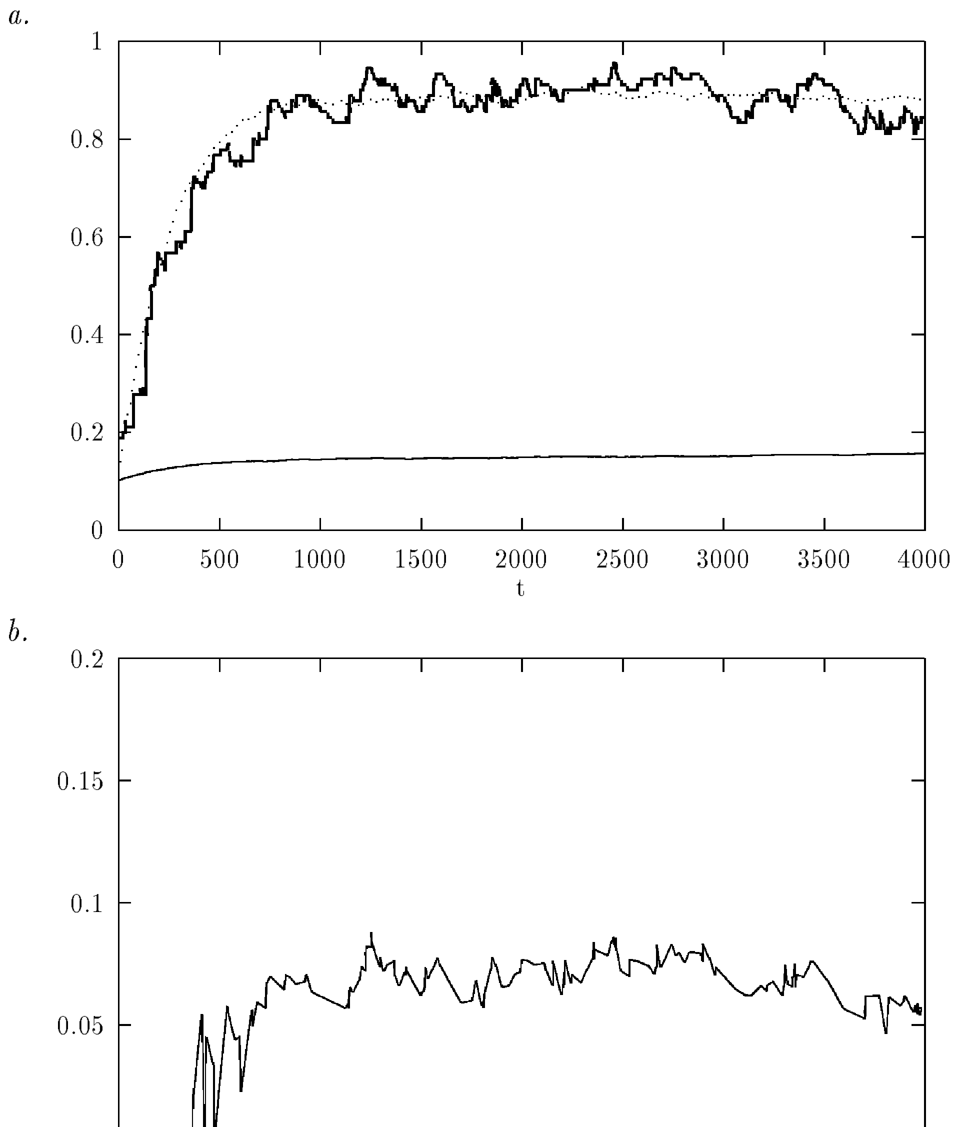

Figure 11: Combined learning and retrieval dynamics – synaptic and neural observables vs time (in units of mean inter-stimulus interval t_c). Total of about 4000 stimuli. (a) Synaptic observables. Bold: Intra-stimulus connectivity (ICC) c_μ for μ=1. Light: connectivity c (initial random connectivity $c(0)$=0.1). Dotted: m_W similarity to Willshaw matrix. (b) Average activity of foreground neurons in attractor after delay, m_+^1, following each stimulus of class number 1. m_0^1, is too small to be distinguished from the axis.

attractors have been silidly learned in a gradual process.

5.2 Varying stimulus statistics and attractor plasticity

Next we have investigated the performance of this learning network when the large scale statisitics of the stimulus stream change after a certain set of attractors has been learned. We have verified that such a change in the input stream leads to an adaptation of the internal representations accounting well for the variations in the input statistics.

The procedure we used is described in Section 2.7. We take p=30 classes divided in three groups C_1, C_2, C_3 of $p_c = 10$ classes each. For $t < T = 2000t_c$ only stimuli belonging to classes of C_1 and C_2 are presented, while for $t > T$ stimuli are restricted to classes of C_2 and C_3. [4]

Fig. 12a presents the evolution of intra-class connectivity c_μ for one prototype in each group. Fig. 12b is the accompanying attractor observable after the delay following a presentation of a stimulus from the corresponding class. One observes that first the classes from groups C_1 and C_2 are learned, with the ICC of C_1 rising faster than that of C_2, due to some fluctuation. It is interesting to observe the corresponding behavior of the attractor recall observable in (b). But after about $500t_c$ the ICC's of the first two groups have been properly learned, i.e have reached high stationary values and the corresponding attractor observables indicate recall on every trial.

At t=$2000t_c$, i.e. after about 2000 presentations of stimuli from the first two groups of classes, the change in the statistics takes place as described above. It provokes a rapid increase of the ICCs for classes belonging to the third group. This increase is accompanied by a decrease of the ICC corresponding to C_1. The latter decreases towards the asymptotic value of the overall connectivity c. This expresses the elimination of the corresponding attractor from memory due to the fact that the corresponding class no longer appears in the stimulus stream. The corresponding attractor exchange can be observed in the lower figure describing the retrieval observables.

Note that the decay of the ICC of C_1 is slower than its initial rise, or of the rise of the ICC's for the new attractor for C_3. In fact, it aproaches its asymptotic value only after thousands of other stimuli have been shown to the network. This feature can be captured analytically[25].

After a transient period an attractor corresponding to the class belonging ot group C_3 appears, while the attractor corresponding to a class that is no longer presented, i.e. from group C_1 is progressively disappearing. The appearance and the elimination of attractors occur when the ICC corresponding to the class is about 0.7. This value is significantly closer to the asymptotic, stationary ICC for the learned classes, which is about 0.8-0.9, than to the asymptotic connectivity c, which is about 0.1-0.2. Thus, while the synaptic observables decay more slowly

[4] It may not be futile to recall that we are dealing with groups of classes of stimuli.

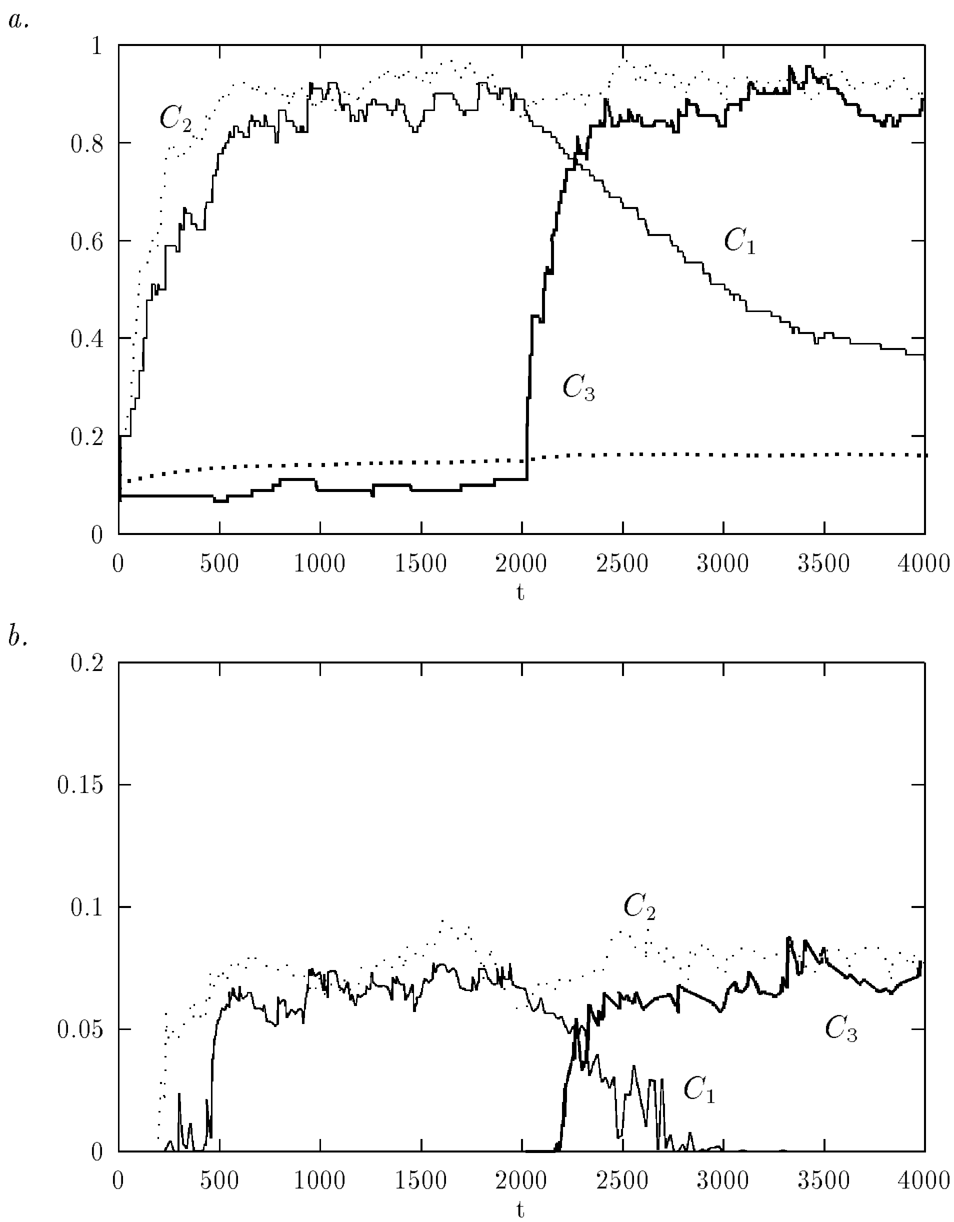

Figure 12: Learning and recall dynamics with time varying stimulus statistics. Statistics change after $2000t_c$ (a total of about 2000 presentations): before, stimuli belong to classes in groups C_1 and C_2; after, to C_2 and C_3. (a) Synaptic observables: intra-class connectivity (ICC) c_μ for one class in each of the three groups (group number indicated next to curve). Bold dotted curve: overall connectivity. (b) Neural attractor observable: Mean activity in the foreground in attractors following delay after presentation of a stimulus of the same three classes (indicated on curves). The mean activity in the corresponding backgrounds is negligible, as in Fig. 11b and is omitted.

than they rise, new attractors appear after about the same amount of training as it takes the old ones to be forgotten.

To visualize the loss of the attractor of the first class of C_1 we present stimuli from the corresponding class rarely. This way, beyond $2000t_c$, there is essentially no learning of the stimuli belonging to the class being forgotten. To gain better statistics, we have presented such stimuli more often, but blocked the corresponding synaptic changes. The role of those stimuli is only to test if a corresponding attractor still exists.

6 Statistics of learned attractors

6.1 Statistics in the input stream

We have seen in Sec. 5 that the system creates attractors which are the internal representations of the classes underlying the stream of stimuli. In order to study how the properties of the set of these attractors depend on the large scale statistics of stimuli in the input stream, we have run simulations with time-independent statistics as described in Sec. 5.1, varying parameters of the input stream.

The dynamics of neurons and synapses is the same as in the double dynamics described in Section 5. The set of prototypes underlying the classes of stimuli can be characterized by three parameters

- Their number p;
- The coding level f or the mean number of active neurons per stimulus $M = fN$;
- The standard deviation, σ, of the number of active neurons in a prototype i.e.

$$\sigma^2 = \frac{1}{p} \sum_{\mu} \left(\sum_{i} \xi_i^{\mu} - M \right)^2$$

Four different procedures for drawing a set of prototypes with p and f fixed have been used.

The first is the distribution in Eq. (13). Its standard deviation is

$$\sigma = \sqrt{f(1-f)N}.$$

For N=200 and f=0.05, we have M=10 and $\sigma \sim 3$. The fluctuations of the number of active neurons are therefore large, with significant influence on the learning process.

The second procedure uses prototypes with a fixed number of active neurons M, i.e. $\sigma = 0$.

The third is an interpolation between the first two. Prototypes are selected as in the first procedure, but those that deviate from the mean (M active neurons) beyond a certain amount are rejected. The resulting distribution has a standard deviation

$$\sigma = \lambda\sqrt{f(1-f)N}$$

where $0 < \lambda < 1$ depends on the rejection criterion.

Last, we have used prototypes which are more inhomogeneous than random ones, i.e. with a larger standard deviation, $\lambda > 1$.

To generate a set of prototypes having mean number of active neurons M and a standard deviation of approximately σ we proceed as follows: for any integer $m \neq M (0 < m \leq p)$ we generate the integers $n(m)$ by rounding the numbers

$$\frac{p}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(m-M)^2}{2\sigma^2}\right)$$

where p is the total number of classes. The first is the number of active neurons in a prototype. The second is the number of prototypes with that number of active neurons. The distribution of coding levels is then approximately Gaussian, with mean M and width σ. The number of prototypes having exactly M active neurons $n(M)$ is given by

$$n(M) = p - \sum_{m \neq M} n(m),$$

to ensure the correct total number of prototypes p.

6.2 Statistics of the attractors

We have seen in Section 5 that after stimuli are presented for a long enough time by one of the procedures described in Sec. 2.7, attractors are created in the network. At any given time, the set of attractors can be characterized by the following observables:

- The number of resulting, learned attractors, or class representatives p_a. Here a learned attractor is an attractor reached by the network following the presentation of a stimulus from one of the classes underlying the input stream. When $p_a = p$ the network has formed attractors corresponding to all the classes in the input stream. How p_a is obtained is explained below.
- The coding level in the learned attractors f_a, or $M_a = f_a N$.
- The standard deviation σ_a of the number of active neurons in the learned attractors.

In the following, we present several examples of the dependence of these attractor observables on the standard deviation σ of the coding level of the stimuli. The network parameters are the same as in section 5, i.e. N=200 neurons, with mean coding level in stimuli f=0.05 (M=10), and learning transition probabilities p_+=0.1

and p_-=0.005. The simulations have been carried out for two different values of the number of classes in the stimulus stream, p=20 and 30. The dependence on f, p_+ and p_- has not been studied.

We expect that the number of attractors that can be learned by the system increases as f decreases (see e.g. ref. [4]). Based on the same theoretical grounds, we also expect that equilibrating the number of potentiations and depressions, i.e. the probabilities satisfy to Eq. (18), optimizes the number of attractors that can be recalled.

For a given set of parameters $\{p, M, \sigma\}$ we generate prototypes with different coding statistics, according to one of the procedures described in Section 6.1, above. From these prototypes stimulus classes are generated, as described in Section 2.7. The stimuli are presented to the network, for learning and retrieval, as in Section 5.

After the transitory period of attractor formation we count the number of active neurons in the attractor in which the network persists following the delay after the presentation of each stimulus. A neuron is considered active if $V > 0.04$ in the attractor. The precise selection value is unimportant since the distribution of activities in an attractor is bimodal: most neurons have zero activities and for the few in the foreground $V_i \sim 0.08$. The attractor activity counting is done in the period $50pt_c < t < 100pt_c$, after the learning observables have reached their asymptotic values (it corresponds to the interval between t=1000 and 2000 in Figs. 11).

The number of learned attractors p_a is defined as the number of classes that have been recognized at **all** presentations. Recognition of the particular class is said to occur if the overlap of the network state with the prototype of the class, $m_+^\mu - m_0^\mu$, is larger than 0.04, at the end of the inter-stimulus interval following the presentation of the stimulus.

For these classes that have been properly learned we perform the statistics on the number of active neurons per attractor in all the presentations. From those we obtain M_a, the average number of active neurons in an attractor, and σ_a, the standard deviation around this number. For each set of parameters we have performed two different runs with different sets of prototypes to check that the results are similar from sample to sample.

The results are presented in Fig. 13. It shows the observed standard deviation of the coding level in attractors σ_a (left) and the number of learned attractors p_a (right) as a function of σ, the spread in the coding of the incoming stimuli, for p=20 and p=30 classes in the input stream. The mean number of active neurons per attractor is found to be about $M_a = 8$. It is lower than the number of active neurons per stimulus. This, we think, is due to the stochasticity of the learning process.

Since the number of attractors which can coexist increases with decreasing coding level, as is the case in Ref. [4] and in preliminary simulations, the above result would imply that the network may learn a larger set of attractors than would have been implied by the coding level in the stimuli.

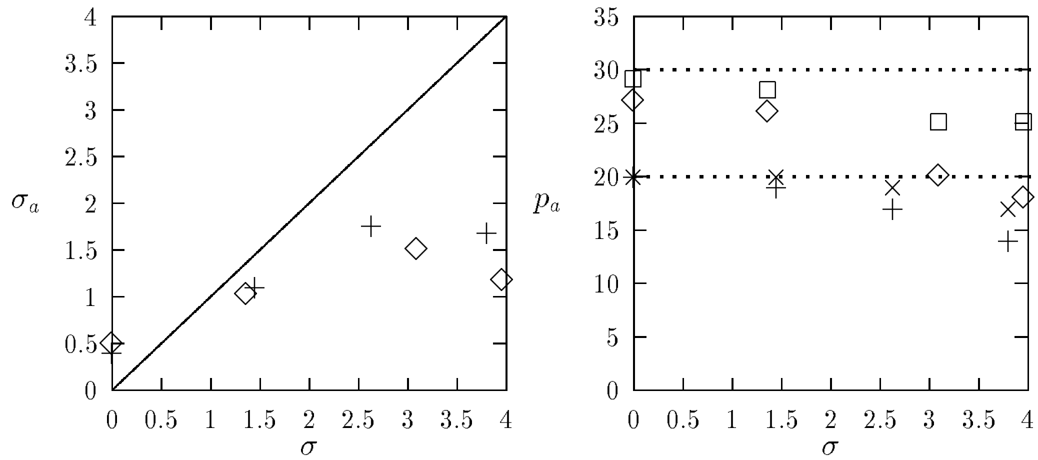

Figure 13: Properties of learned attractors vs stimulus coding statistics. Left: standard deviation, σ_a, of coding level in attractors vs σ in stimuli. Right: number of learned attractors vs σ in stimuli. $\Diamond$: $p = 30$, classes recognized at all presentations. $\Box$: $p = 30$, classes recognized at 90 % of presentations. $+$: $p = 20$, classes recognized at all presentations. $\times$: $p = 20$, classes recognized at 90 % of presentations.

Another interesting finding is that the fluctuations in the number of active neurons per attractor are reduced relative to the corresponding fluctuations among the class prototypes, Fig. 13(left). This is a consequence of the fact that classes that activate a small number of neurons are more difficult to learn, because one needs a critical number of active neurons (close to fN) for an attractor to be sustained by the network. On the other hand, classes activating a large number of neurons will form an attractor with less active neurons to avoid the generation of excessive inhibition. The learned set of attractors is more homogeneous than the corresponding set of prototypes, since learning discards stimuli that activate too few neurons and smoothes out the others. Coding level fluctuations have significant effects on the performance of Willshaw type networks even if the number of neurons is very large[26, 27]. The fact that learning a set of attractors reduces these fluctuations may improve the performance of the network.

The dependence of the number of learned attractors, p_a, on the standard deviation of the coding level in the stimuli σ is illustrated in Fig. 13 (right). It shows that when σ=0 (all prototypes have an equal number of active neurons), the network can perfectly learn a set of p=20 classes, while from a set of p=30 classes it has perfectly learned 27 of them.

However if we use a 90% success criterion instead of the 100% one for class recognition, we find than 29 classes have been learned. The remaining class, it turns out, has been presented an unusually low number of times during learning, due to the stochasticity of the presentations, and the corresponding attractor has not been permanently formed in the synaptic structure.

This figure exposes also the influence of the parameter σ on the number of learned attractors. As σ increases less classes can be learned perfectly, because classes whose prototype deviate too much from the mean coding level will be harder to learn.

7 Conclusions and outlook

We have studied a model of an autonomously learning attractor neural network, in the framework of a project of an implementation of a larger network including an input and output parts [2]. The simulated network, which corresponds to the central part, has been shown to be able to learn representations of the stimuli shown on the input side, and to be quite robust to many presentation procedures and/or noise in the presentations. The internal representations in the network are plastic, and adapt in the case of a macroscopic change in the input statistics. The effort has been to build the simplest possible network with analog neural variables.

In the present network a single inhibitory neuron can control the global activity in the excitatory network and prevent catastrophic events like the percolation of the activity to the whole network. It also serves to 'turn off' the activity of all excitatory elements when an unfamiliar stimulus is presented.

Based on exeprience gained in the study of Ref. [2] we have also here used the fact that synaptic elements which have as few as two stable efficacies, and with a local hebbian dynamics, make possible the formation of internal representations of the stream of stimuli in the network. For the sake of speeding up the simulations we have replaced the analog synaptic dynamics by a related stochastic process between synapses of two states. A synaptic matrix with this dynamics can adapt to changes in the statistics of the input stream, and 'forget' internal representations corresponding to classes of stimuli eliminated from the input stream.

The network keeps the essential features of associative memories, with the advantage of continual learning. It can be an ideal testground for the study of the formation of internal representations through learning in a variety of contexts: from neuro-physiology to data acquisition. A material device implementing a learning attractor network with discrete neurons has already been implemented [2]. The present model is being implemented in silicon.

We turn now to several issues that have not been adressed in the present work and should be considered in the near future. In certain situations there may be significant differences between the analog synaptic dynamics assumed to underlie learning and the stochastic process actually implemented and tested here. To test for those differences requires very large computational power because of the difference in time scales of neural and synaptic dynamics. An electronically implemented network of the type discussed here is an ideal ground for such tests, as well as for testing many different speculations involving both neural and synaptic temporal variation.

In this study the input has been considered to show binary configurations to the network. More realistic synaptic inputs would be analog, as would be the case for an input 'retina' connected to the attractor network with receptive fields, such as e.g. Ref. [7]. In this case the distribution of the neural activities when the network is in an attractor is likely not to be bimodal like in the present case. How this situation affects learning of representations of the stimuli shown on the 'retina' is of great interest.

In the present network learning occurs in all possible synapses, and the connectivity is determined by learning. A more realistic situation would have an initial connectivity matrix determining for every neuron who are its synaptic neighbors, and to have learning dynamics occur only on the links present.

In the present study we have considered learning representatives of uncorrelated classes only. It is a clear item on the agenda to study the consequences of the learning dynamics for a stream of stimuli containing correlated classes.

Another issue that can be considered is how the synaptic dynamics makes possible the learning of temporal correlations between stimuli [16, 5]. We expect that if the short term synaptic changes are proportional not to the instantaneous pre and post synaptic activities, but rather to some average of these activities over some temporal window, these temporal correlations are naturally converted in spatial correlations between the internal representations. The spatial correlations should increase with the size of the averaging window. Another possible scenario for the conversion of temporal correlations between stimuli in spatial ones is provided by a careful examination of Fig. 6. If the network is initially in an attractor, and a different stimulus is shown, there is a short interval — of the order of the neuronal integration constant — in which both neurons participating in the old attractor and neurons activated by the new stimulus are significantly active. In this interval LTP could occur in the synapses connecting the two populations of neurons. This, in turn, could have a significant effect on the synaptic matrix if these two stimuli are repeatedly shown together [28]. This did not occur in the present study, for two reasons. First, the threshold for LTP was too high for this type of transitions to occur. Second, the average presentation time of a stimulus was much longer than the neuron integration time, and thus the probability of any transition, even in the case of threshold crossing, was negligible. But in principle one can lower the threshold for LTP and shorten the presentation time so that these transitions can occur and have a significant effect in the statistics of the attractors. This question deserves further study.

Acknowledgements

We wish to thank Stefano Fusi for many discussions and comments on a previous version of the manuscript. This study was done in the context of a project to develop an electronic implementation of a learning attractor neural network, and has been supported in part by a grant from the Consiglio nazionale della Ricerca, Italy. NB was supported by a fellowship of programme Cognisciences of CNRS, France.

References

[1] Badoni D, Riccardi R and Salina G 1992 Learning attractor neural network: the electronic implementation, *Int. J. of Neural Systems*, **3** 13 (Proceedings of Neural Networks: from Biology to High Energy Physics, Isola d'Elba)

[2] Badoni D, Bertazzoni S, Buglioni S, Salina G, Amit DJ and Fusi S 1995 Electronic implementation of a learning attractor neural network, to be published in *Network*.

[3] Amit DJ and Fusi S 1992 Constraints on learning in dynamic synapses, *Network*, **3** 443

[4] Amit DJ and Fusi S 1994 Learning in neural networks with material synapses, *Neural Computation*, **6**

[5] Amit DJ, Brunel N and Tsodyks MV 1994 Correlations of cortical Hebbian reverberations: theory *vs* experiment, *J. Neurosci.*, **14** 6445

[6] Amit DJ and Brunel N 1995 Global spontaneous activity and local structured (learned) delay activity in cortex, in preparation

[7] Amit DJ and Brunel N 1993 Adequate input for learning in attractor neural networks, *Network* **4** 177

[8] Amit DJ and Tsodyks MV 1991 Quantitative study of attractor neural networks retrieving at low spike rates I, *Network* **2** 259

[9] Amit DJ and Tsodyks MV 1991 Quantitative study of attractor neural networks retrieving at low spike rates II, *Network* **2** 275

[10] Amit DJ and Tsodyks MV 1992 Effective neurons and attractor neural network in cortical environment, *Network* **3** 121

[11] VanVreeswijk C and Hasselmo ME 1995 Self-sustained firing in a simple model with excitatory and inhibitory neurons, in preparation.

[12] Marr D 1970 A theory for cerebral neocortex, *Proc. Roy. Soc. Lond. B. 176* 161

[13] Traub RD and Miles R 1991 *Neuronal Networks of the hippocampus* (Cambridge University Press, Cambridge)

[14] Amit DJ 1989 *Modeling Brain Function* (Cambridge University Press, NY)

[15] Wilson FAW, Scalaidhe SPO and Goldman-Rakic PS 1993 Dissociation of Object and Spatial Processing Domains in Primate Prefrontal Cortex, *Science*, **260** 1955

[16] Miyashita Y 1988 Neuronal correlate of visual associative long-term memory in the primate temporal cortex, *Nature* **335** 817

[17] Dong D and Hopfield DJ 1992 Dynamic properties of neural networks with adapting synapses *Network* **3** 267

[18] Willshaw D, Buneman O P & Longuet-Higgins H 1969 Non-holographic associative memory, *Nature*, **222** 960

[19] Nadal JP and Toulouse G 1989 Information storage in sparsely coded memory nets *Network* **1** 61

[20] Miyashita Y and Chang HS 1988 Neuronal correlate of pictorial short-term memory in the primate temporal cortex, *Nature*, **331** 68

[21] Sakai K and Miyashita Y 1991 Neural organization for the long-term memory of paired associates, *Nature*, **354** 152.

[22] Amit DJ 1992 In defence of single electrode recording, *Network* **3** 385

[23] Frégnac Y, Smith D and Friedlander MJ 1990 Postsynaptic membrane potential regulates synaptic potentiation and depression in visual cortical neurons, *Society for Neuroscience Abstracts* **16** 798

[24] Stanton PK and Sejnowsky TJ 1989 Associative long-term depression in the hippocampus induced by Hebbian covariance, *Nature* **339** 215

[25] Brunel N and Fusi S 1995 in preparation.

[26] Nadal JP 1991 Associative memory: on the (puzzling) sparse coding limit, *J. Phys. A* **24** 1093

[27] Brunel N 1994 Storage capacity of neural networks: effect of the fluctuations of the number of active neurons per memory, *J. Phys. A*, **27** 4783.

[28] Brunel N 1995 Stochastic learning of temporal correlations in attractor neural networks, in preparation.