Dipartimento di Fisica
Universitá di Roma "La Sapienza"
I.N.F.N. - Sezione di Roma

ROME1-1104/95
June 12, 1995

ROME-1-1104-95
Fermilab Library
0 1160 0073976 7
FERMILAB
JUL
LIBRARY
34

*N.Brunel*

# HEBBIAN LEARNING OF CONTEXT IN RECURRENT NEURAL NETWORKS

# Hebbian learning of context in recurrent neural networks

**Nicolas Brunel**
**INFN, Sezione di Roma, Istituto di Fisica**
**Università di Roma I 'La Sapienza'**
**P.le Aldo Moro 2, 00185 Roma Italy**

**Abstract**

Single electrode recordings in inferotemporal cortex of monkeys during delayed visual memory tasks provide evidence for attractor dynamics in the observed region. The persistent elevated delay activities could be internal representations of features of the learned visual stimuli shown to the monkey during training. When uncorrelated stimuli are presented during training in a fixed sequence, these experiments display significant correlations between the internal representations. Recently a simple model of attractor neural network has reproduced quantitatively the measured correlations. An underlying assumption of the model is that the synaptic matrix formed during the training phase contains in its efficacies information about the contiguity of persistent stimuli in the training sequence. We present here a simple unsupervised learning dynamics which produces such a synaptic matrix if sequences of stimuli are repeatedly presented to the network at fixed order. The resulting matrix is then shown to convert temporal correlations during training into spatial correlations between attractors. The scenario is that, in presence of selective delay activity, at the presentation of each stimulus, the activity distribution in the neural assembly contains information both of the current stimulus as well as of the previous one (carried by the attractor). Thus the recurrent synaptic matrix can code not only for each of the stimuli which were presented to the network, but also for their context. We combine the idea that for learning to be effective synaptic modification should be stochastic, with the fact that attractors provide learnable information about two consecutive stimuli. We calculate explicitly the probability distribution of synaptic efficacies as a function of training protocol, i.e. the order in which stimuli are presented to the network. We then solve for the dynamics of a network composed of integrate-and-fire excitatory and inhibitory neurons with a matrix of synaptic collaterals resulting from the learning dynamics.

The network has a stable spontaneous activity, and stable delay activity develops after a critical learning stage. The availability of a learning dynamics makes possible a number of experimental predictions for the dependence of the delay activity distributions and the correlations between them, on the learning stage and the learning protocol. In particular it makes specific predictions for pair-associates delay experiments.

# 1 Introduction

## 1.1 Correlated delay activities – experiment and theory

In the last twenty years there has been a wealth of evidence for the existence of local reverberations of cell assemblies in inferotemporal cortex (Fuster and Jervey 1981, Miyashita and Chang 1988, Miyashita 1988, Sakai and Miyashita 1991, Tanaka 1992), prefrontal cortex (Fuster 1973, Niki 1974, Goldman-Rakic 1987, Wilson et al 1993), and other areas of primates during delayed visual memory tasks (for a review see Fuster 1995). Together with experimental data, models have been proposed to account for the persistent delay activities (Dehaene and Changeux 1989, Zipser et al 1993, Griniasty et al 1992), in which excitatory synapses store the information about the visual stimuli. The experiments of Miyashita (1991) on the activity in IT cortex of monkeys trained to perform a DMS task have disclosed significant correlations in the persistent delay activities following the presentation of uncorrelated stimuli, when those are presented during training in a fixed sequence.

Theoretical studies (Griniasty et al 1992, Amit et al 1994, Brunel 1994) have demonstrated that attractor neural networks which embed in their synaptic structure information about contiguous stimuli learned in a sequence, have correlated delay activities even though the learned stimuli are uncorrelated. It may be worth pointing out that when stimuli arrive at IT they may be uncorrelated because they have been so prepared, or because they have been decorrelated on the way (Barlow 1961, Linsker 1989, Atick 1992). In the model networks, the delay activity provoked in the neural assembly by the presentation of a given learned stimulus is correlated with the delay activity corresponding to other stimuli until a separation of several stimuli in the training sequence, despite the fact that the synaptic matrix connects only consecutive stimuli in the sequence. The appearance of such correlations between the different delay activities is a transcription, during the learning process, of temporal correlations in the training information, into spatial (activity distribution) correlations of the internal representations of the different stimuli. The network has therefore a memory of the context of the presented stimuli. Some cognitive implications of this context sensitivity have been outlined in (Amit 1995).

The model simulated by Amit et al (1994) consists of a network of integrate-and-fire neurons represented by their current to spike rate transduction function (Amit and Tsodyks 1991). Such neurons are taken to represent the excitatory neurons of the network, the pyramidal cells. It is in the synaptic matrix connecting these neurons that learning is manifested. The synaptic matrix, representing the training process, is constructed to represent the inclusion of the information about the contiguity of patterns in the training sequence, as in (Griniasty et al 1992). Inhibition is taken to have fixed synapses and its role is to react in proportion to the mean level of activity in the excitatory network, so as to control the overall activity in the network. The delay activities are investigated by presenting to the neural

module one of the uncorrelated stimuli as a set of afferent currents into a subset of the excitatory neurons. These currents are removed after a short time and the network is allowed to follow the dynamics as governed by the feedback represented in the matrix of synaptic collaterals. Eventually, the network arrives at a stationary distribution of spike rates. This is the delay activity distribution corresponding to the stimulus which excited the network. Simulations of the model (Amit et al 1994) are in quantitative agreement with the experimental data of Miyashita (1991).

The dynamics of the model has been solved analytically in simplified conditions (Brunel 1994). This makes possible the explicit calculation of the correlations between the internal representations, as a function of the parameters of the model. The main parameters controlling these correlations are the strength of the inclusion of the contiguity between stimuli in the synaptic matrix, relative to the strength of the inclusion of the stimuli themselves, and the balance between recurrent excitatory and inhibitory synaptic efficacies. The analysis deduces the mean fraction of neurons activated by a given stimulus (coding level, or sparseness) in the observed region, from the experimental data of (Miyashita 1991). This in turn makes possible the calculation of the correlation coefficients, which are again in quantitative agreement with all the available experimental data (see Fig. 9 of Brunel 1994), and the simulations of Amit et al (1994).

These previous studies (Griniasty et al 1992, Amit et al 1994, Brunel 1994) used a fixed pre-arranged synaptic matrix. In (Amit et al 1994, Brunel 1994) the matrix was chosen to be similar to the Willshaw matrix (Willshaw et al 1969), with a limited number of synaptic states. Memory is coded exclusively in the excitatory-to-excitatory synapses. An important result (Amit et al 1994) is that the correlations are rather insensitive to the particular matrix chosen, provided it is Hebbian and that it includes the memory of the contiguity between stimuli.

What is missing is a plausible dynamic learning process leading to a synaptic matrix which incorporates information of the temporal context of the stimuli shown to the network. One way of implementing learning dynamics is to allow for each synaptic efficacy a limited number of stable values (Amit and Fusi 1994). Learning, which may be analog on the short term, becomes a walk between the discrete stable efficacies in the long term. To make such learning efficient, transitions between the different states, provoked in a Hebbian way during the presentation of a stimulus by the activity of pre and post synaptic neurons, should be stochastic. Such dynamics has been simulated (Amit and Brunel 1995a) and analyzed (Amit and Fusi 1994). A synaptic matrix endowed with such a dynamics is able to learn internal representations of the classes of stimuli shown to the network. However the stochastic process studied by Amit and Fusi (1994) precludes the possibility of learning any temporal correlations between stimuli.

## 1.2 The present work

In the following we first discuss a possible scenario for learning in presence of delay activity which naturally leads to the inclusion of temporal correlations between stimuli in the synaptic matrix. The scenario is that first uncorrelated attractors are formed. An attractor then carries information from the stimulus that provoked it until the presentation of the next stimulus. This information allows for a simple synaptic mechanism to store the memory of the context of any stimulus. We study the case of a finite set of stimuli which are repeatedly shown to the network. In the simplified case in which every excitatory neuron in the network is activated by at most one stimulus (Brunel 1994), it is possible to calculate explicitly the probability distribution of every synaptic efficacy as a function of the learning procedure. If stimuli are shown repeatedly in a fixed order during learning, the resulting synaptic matrix is similar to the fixed matrix used in (Amit et al 1994, Brunel 1994). Given the synaptic matrix we solve for the neural dynamics of the attractor network as in (Brunel 1994), when one of the stimuli is presented. The generic features of such a learning process will be discussed elsewhere (Brunel and Fusi 1995).

The network we study is composed of a large number of excitatory and inhibitory integrate-and-fire neurons, described by the statistics of their afferent currents and their spike emission rates. The network represents a local module, similar to a cortical column, embedded in a much larger sea of neurons (the entire cortex). The module can be distinguished from the global network by two features: the high local excitatory connectivity and the range of inhibitory interactions (Braitenberg and Schuz 1991). Such a network has a stable state of low activity in which all neurons have a spontaneous activity of the order of 1-5 spikes per second in a plausible region of parameters (Amit and Brunel 1995b). Furthermore, when learning occurs in the local module, and the synaptic modifications are strong enough, a set of attractors correlated with the stimuli presented to the network develops. In each attractor a small subset of the excitatory neurons — the neurons which are activated by a particular stimulus — have elevated delay activities, of the order of 20-40 spikes per second. We choose to study both learning and retrieval dynamics in this network since the activity in its attractors is roughly in agreement with recorded data during DMS experiments in both inferotemporal and prefrontal cortex.

When learning occurs in the present network, upon repeated presentation of stimuli, uncorrelated attractors are initially formed. These attractors make possible the inclusion of temporal correlations between stimuli in the synaptic matrix. This in turn provokes significant correlations in the delay activities corresponding to stimuli which have been shown repeatedly contiguously to the network. Therefore the correlations between the internal representations of different stimuli reflect their context.

Using a plausible learning process one reproduces the results found in (Amit

et al 1994, Brunel 1994), which are in good agreement with experimental data (Miyashita 1991). This is not surprising since the synaptic matrix resulting from many presentations of the stimuli is quite similar to the matrix that was postulated in (Amit et al 1994, Brunel 1994). One essential novelty is that the entire phenomenon takes place in presence of stable spontaneous activity. The advantage of using the more realistic neural model of Amit and Brunel (1995b) is that neurons have both spontaneous and selective activity roughly in the range of the recorded data.

The analysis allows to predict:

- The evolution of the delay activities and of the correlations between the internal representations during training, for a fixed training procedure;
- The dependence of the correlations on the training procedure.

The predictions of the theory are accessible to experiments as in (Miyashita and Chang 1988, Miyashita 1991, Sakai and Miyashita 1991). We focus the analysis on two particular cases.

- Training with stimuli in a fixed sequence, as in (Miyashita 1991).
- Training with associated pairs, as in (Sakai and Miyashita 1991): a set of stimuli is divided into pairs. Stimuli in each pair are presented in fixed order. Pairs are presented at random.

We also show how it is possible to deal with intermediate cases, as when the sequence of stimuli is interspersed with random items.

The paper is organized as follows. In section 2 we present in detail the model network and its elements. In the following section we present a simple scenario of synaptic dynamics which incorporates both associative LTP and LTD. Then we describe a typical protocol of a visual memory experiment in which a delay period always follows the presentation of a stimulus. We show that in this situation the analog synaptic dynamics reduces to a stochastic process acting on a two state synapse. We then study in detail which kind of synaptic transitions may occur, depending on whether there is selective delay activity following the presentation of a stimulus or not. In section 4 we study the situation of a small set of stimuli repeatedly shown to the network. In this case we calculate explicitly the probability distribution of the synaptic efficacies of the network as a function of the learning stage and of the learning protocol. Then, in section 5, we study the network dynamics and show the influence of the synaptic dynamics on the delay activity which is stabilized by the network after the presentation of a learned stimulus. This allows to study the structure of the delay activity distributions as a function of the learning stage and the learning protocol.

## 2 The model neurons

Each neuron in the network receives three types of inputs: from recurrent (collateral) excitatory connections from other neurons in the same network; from inhibitory neurons inside the network; from excitatory neurons in other, unspecified, areas. The collateral connectivity in the network has no geometric structure: a neuron has equal probability (about 0.1) of having a synapse on any other neuron.

Both excitatory and inhibitory neurons are leaky integrate-and-fire neurons described by the statistics of their input currents, which determines their firing rates (Amit and Brunel 1995b). Each type of neuron is characterized by a threshold $\theta_\alpha$, a post-spike hyperpolarization $H_\alpha$, an integration time constant $\tau_\alpha$, with $\alpha = E, I$ indicating whether the neuron is excitatory or inhibitory, respectively. A neuron $i$ of type $\alpha$ receives a large number of afferent spikes per integration time (Amit and Brunel 1995b), and hence a Gaussian white noise input current of mean $I_i^\alpha$ and standard deviation $\sigma_i^\alpha$, through $C_\alpha$ synaptic contacts, which are divided in $C_{\alpha E}$ excitatory synapses and $C_{\alpha I}$ inhibitory ones.

The synapses in the network are of four types, depending on all the possible types of pre and post synaptic neurons. For each synaptic type the efficacies $J_{ij}$ ($i$ and $j$ denote the post and pre synaptic neuron, respectively) are drawn randomly from the distribution $P_{\alpha\beta}(J)$ ($\alpha$ and $\beta$ denote the type of post and pre synaptic neuron, respectively). $P_{\alpha\beta}$ has mean $J_{\alpha\beta}$ and standard deviation $J_{\alpha\beta}\Delta$, where $\Delta$ represents the variability in the synaptic amplitude. A fraction $x_\alpha$ of the excitatory connections on a neuron of type $\alpha$ arrive from outside the network. The excitatory to excitatory connections are plastic: the distribution $P_{EE}(J)$ specifies the distribution of excitatory to excitatory links before the learning stage. As we will see later learning will modify this synaptic distribution.

The spike rate of excitatory neuron $i$ is $\nu_i^E$. The rate of inhibitory neuron $i$ $\nu_i^I$. The input currents from outside the column are described by a white noise with mean $I_i^{ext}$ and standard deviation $\sigma_i^{ext}$. This input currents are provoked, in absence of a stimulus, by the background activity outside of the network. In presence of a stimulus, the input currents are the sum of the background input and of the input provoked by that stimulus.

We assume that the correlations between the spike emission times of different neurons in the network do not affect significantly their spike rates. Thus we consider the spike emission processes of different neurons in the network as uncorrelated. In this case the mean and variance of the input current to a neuron in the module are the sum of three independent contributions, coming from external excitatory, recurrent excitatory, and inhibitory currents (see Amit and Brunel 1995b)

$$I_i^\alpha = I_i^{ext} + \tau_\alpha \sum_{j\in E} J_{ij}^{\alpha E}\nu_j^E - \tau_\alpha \sum_{j\in I} J_{ij}^{\alpha I}\nu_j^I \tag{1}$$

and

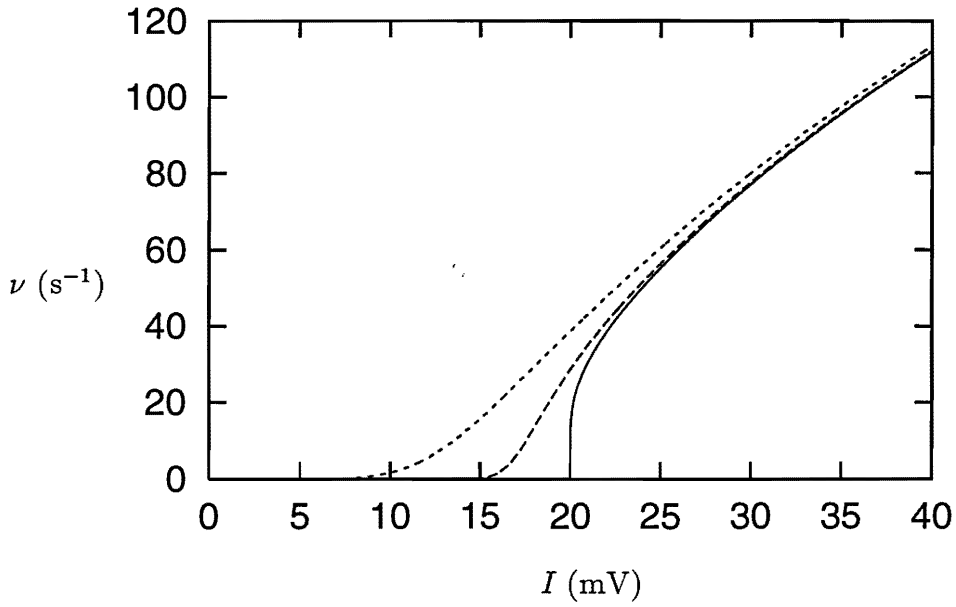


Figure 1: Current to frequency transduction function $\nu = \phi(I, \sigma)$ for $\theta$=20mV, $H$=0, $\tau$=10 ms, $\tau_0$= 2 ms and three values for the amplitude of the fluctuations of the currents $\sigma$= 0 (full line), 2 mV (dashed line) and 5 mV (dotted line).

$$(\sigma_i^\alpha)^2 = (\sigma_i^{ext})^2 + \tau_\alpha \sum_{j\in E} \left(J_{ij}^{\alpha E}\right)^2 \nu_j^E + \tau_\alpha \sum_{j\in I} \left(J_{ij}^{\alpha I}\right)^2 \nu_j^I. \tag{2}$$

These currents are integrated by the membrane depolarization at the soma with a time constant $\tau_\alpha$. The firing rate of neuron $i$ of type $\alpha$ is given by

$$\nu_i^\alpha = \phi_\alpha(I_i^\alpha, \sigma_i^\alpha)$$

where

$$\phi_\alpha(I, \sigma) = \left(\tau_0 + \tau_\alpha \int_{\frac{H_\alpha - I}{\sigma}}^{\frac{\theta_\alpha - I}{\sigma}} du \sqrt{\pi} \exp(u^2)[1 + \mathrm{erf}(u)]\right)^{-1} \tag{3}$$

is the transduction function(Ricciardi 1977), which depends on the absolute refractory period $\tau_0$, the threshold $\theta_\alpha$ and post-spike hyperpolarization, or reset potential, $H_\alpha$. The function $\phi$ is plotted as a function of $I$ for three different values of $\sigma$ in Fig. 1. It shows that the fluctuations of the currents have a significant effect on the spike rates when the average current depolarizes the neuron below threshold. Note that the precise form of the transduction function, Eq. 3, is not necessary for the qualitative features of the behaviour of the network.

In the following we take: $\theta_E = \theta_I = 20$ mV above the resting potential; $H_E = H_I = 0$; $\tau_E = 10$ ms; $\tau_I = 2$ ms; $\tau_0 = 2$ ms.

The connectivity parameters are: $x_E = x_I = 0.5$; $C_{EE} = C_{IE} = 20000$; $C_{EI} = C_{II} = 2000$. The average synaptic efficacies are expressed by the amplitude of the

(excitatory or inhibitory) postsynaptic potential provoked by a spike, and thus in units of the potential: $J_{EE} = 0.04$ mV; $J_{IE} = J_{II} = 0.14$ mV; $J_{EI} = 0.05$ mV. The synaptic variability is taken to be $\Delta = 1$. The synaptic external input has mean $I^{ext} = 11$ mV and RMS $\sigma^{ext} = 0.9$ mV into excitatory neurons, and $I^{ext} =$ 8.6 mV and RMS 1.6 mV into inhibitory neurons. These currents correspond to the activation of all the excitatory synapses coming from outside the network at a background rate of 3 $s^{-1}$. For these parameters the network has a stable state of spontaneous activity in which excitatory neurons emit about 3 spikes per second, while inhibitory ones emit 4.2 spikes per second.

Note that this set of parameters is in a biologically plausible region (Braitenberg and Schuz 1991, Komatsu et al 1988, Mason et al 1991). The excitatory to excitatory synaptic efficacy is slightly smaller than the reported range of unitary EPSPs in neocortex and hippocampus, but we have here a neuron that sums linearly its inputs. When the input is nonlinear a larger number of EPSPs are necessary to reach threshold than for a linear input, so the *effective* synaptic efficacy would be smaller than the reported values in the case of a large number of inputs. In fact, the qualitative features to be discussed are fairly robust to small changes in the synaptic efficacies. If the inhibitory efficacies are weakened too much relative to the excitatory efficacies, the spontaneous activity state becomes unstable (Amit and Brunel 1995b).

# 3 Learning dynamics

## 3.1 Analog short term synaptic dynamics

Excitatory-to-excitatory synapses in the network are plastic. Hebbian learning is modelled by a synaptic dynamics which incorporates both associative long term potentiation (LTP) and long term depression (LTD) (Amit and Brunel 1995a):

$$\tau_c \dot{J}_{ij}(t) = -J_{ij}(t) + c_{ij}(t) + (J_1 - J_0)\Theta\left(J_{ij}(t) - w_{ij}(t)\right) + J_0. \quad (4)$$

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

$$c_{ij}(t) = \lambda_+ \nu_i(t)\nu_j(t) - \lambda_-[\nu_i(t) + \nu_j(t)] \quad (5)$$

$\lambda_{+,-}$ 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 very low $c_{ij}$ is negligible.

The last term on the right hand side of Eq. 4 is the 'refresh' mechanism discussed in detail in Badoni et al (1995). It represents one way of preventing the loss of

memory due to the decay of the integrator when no source is present. If at any given moment the source $c_{ij}(t)$ exceeds the fluctuating threshold $w_{ij}(t)$, a refresh source turns on to drive the synapse to the high value $J_1$. If later the source vanishes this synaptic value will remain above its threshold and the efficacy $J_1$ 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, and in the absence of a source that synapse decays to $J_0$. 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. This type of learning is realistic in the sense that it can be (and has been) implemented in a material device (Badoni et al 1995).

As a consequence, in absence of the source term each synapse has two asymptotically stable values, $J_0$ and $J_1$. We further assume that the fluctuations of the threshold are limited to an interval $[J_0 + \theta_+, J_1 - \theta_-]$. The fluctuating threshold therefore defines a potentiation threshold $\theta_+$ such that if the synaptic value is initially low, there is a finite transition probability $J_0 \to J_1$ when the source $c_{ij} > \theta_+$, and a depression threshold $\theta_-$ such that if the synaptic value is initially high, there is a finite transition probability $J_1 \to J_0$ when $c_{ij} < -\theta_-$. These thresholds are such that $J_0 < J_0 + \theta_+ < J_1 - \theta_- < J_1$. We illustrate in Fig. 2 two examples of the evolution of the synaptic efficacy upon presentation of a stimulus. In both cases the synaptic efficacy is initially at $J_0$ and the source term $c_{ij}$ is higher than the threshold $\theta_+$. In the upper figure the synaptic efficacy does not cross the fluctuating threshold and decays to its low stable value after the stimulus is removed. In the lower figure the synaptic efficacy crosses the threshold and is driven to the high state $J_1$, which is stable in absence of a stimulus.

## 3.2 Learning protocol and external currents

The schematic learning protocol we model is as follows. The stimuli shown to the network are labelled by $\mu = 1, \ldots, p$. During the presentation of stimulus $\mu$, the mean external current received by an excitatory neuron $i$ is incremented selectively by $I_{sel}\eta_i^\mu$, where $\eta_i^\mu = 1, 0$ is the symbolic indication of whether cell $i$ is activated by stimulus $\mu$ or not. In absence of a stimulus the excitatory afferent is just the spontaneous noise. Inhibitory neurons are not activated by the stimulus. The presentation of a stimulus is followed by a delay period of length $t_d$, in which the selective part of the current is removed. Therefore, a typical experiment can be schematized by Fig. 3 in which presentation and delay intervals are kept fixed. The duration of each presentation $t_p$ is taken to be much longer than the neuronal time constants $\tau_{E,I}$. Thus $t_p \gg 10$ ms.

Note that in a delayed match to sample (DMS) experiment the sequence of stimuli is an alternate sequence of sample and match stimuli. The match stimulus is typically taken to be equal to the sample stimulus with 50% probability, and

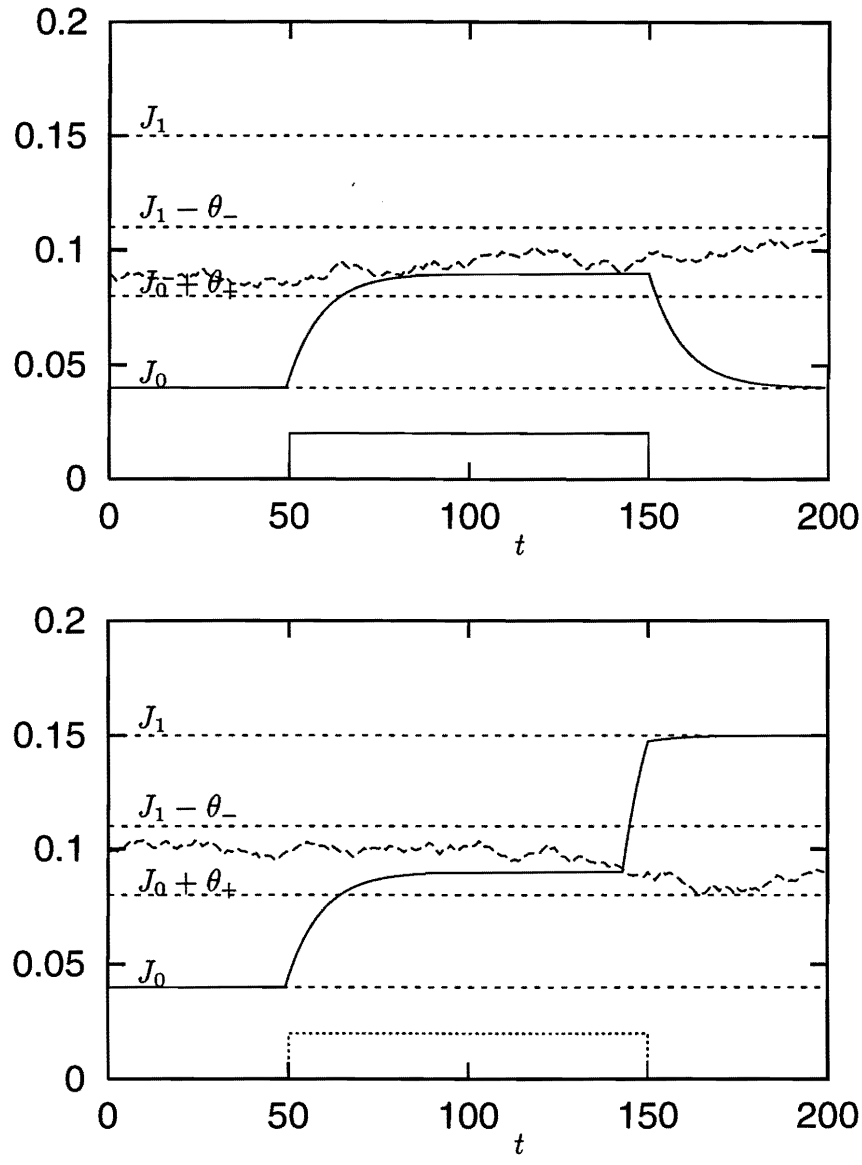


Figure 2: Analog synaptic dynamics. Synaptic efficacy (full line) initially at $J_0$. An external stimulus imposes $c_{ij} > \theta_+$ during the interval $50 < t < 150$. In the upper figure, the synapse does not cross the fluctuating threshold (dashed line) and remains in its low state $J_0$. In the lower figure, the synapse crosses the fluctuating threshold and makes a transition towards the high state $J_1$. Parameters: $J_0 = 0.04$ mV; $J_1 = 0.15$ mV; $\theta_+ = 0.04$ mV; $\theta_- = 0.04$ mV.

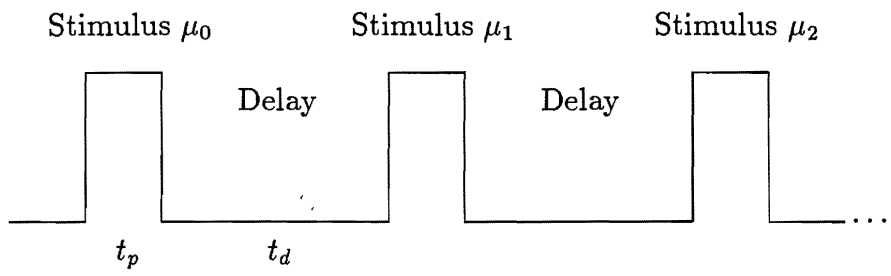


Figure 3: Typical learning protocol in a 'visual memory' experiment. Stimuli are presented in a sequence, with a delay between two successive presentations. The line represents schematically the level of external currents to the local network.

another randomly chosen stimulus otherwise. The learning protocol specifies how the sequence of *sample* stimuli is presented (see below).

To simplify the discussion we suppose that when stimulus $\mu$ is shown, the activated excitatory neurons go rapidly to a steady state rate $\nu_i$:

$$\nu_i = (V - \nu_s)\eta_i^\mu + \nu_s$$

where $\nu_s$ is the spontaneous rate of excitatory neurons, during presentation of the stimulus. When neuron $i$ is activated by a stimulus it goes to a high activity state $V \gg \nu_s$ while if it is not activated it stays at spontaneous activity levels. When the stimulus is removed two possibilities may occur (Amit and Brunel 1995a):

- the stimulus is unfamiliar: the network goes rapidly into its uniform, unstructured, spontaneous activity state,

$$\nu_i = \nu_s$$

- the stimulus is familiar: the activity of neurons which are activated during the presentation of the stimulus persists during the delay period, but with lower rates than during the presentation

$$\nu_i = (v - \nu_s)\eta_i^\mu + \nu_s$$

  where $V > v > \nu_s$.

Following the delay period, when the next stimulus is presented, there is a short interval in which both neurons active in the delay period and neurons activated by the next stimulus will be active. Later inhibition turns off the activity of the neuron which participated in the attractor in the delay period, leaving active only

those neurons which are tagged by the new stimulus (Amit and Brunel 1995a). This transient interval is assumed to be short compared to the presentation time. It will be typically of the order of the integration time $\tau_E$ of an excitatory neuron.

We further assume that the delay period is much longer than the synaptic integration time constant $\tau_c$. In this case, in absence of delay activity, at the end of the delay period all synapses in the network will have decayed to their asymptotic values, i.e. $J_0$ or $J_1$.

## 3.3 Synaptic transitions — no delay activity prior to presentation

We first consider the case in which there was no delay activity before the presentation of the stimulus. When a stimulus is presented, one of eight situations may occur at a given synaptic site $J_{ij}$. For each of the two possible stable values of the synapse ($J_0$, $J_1$) there are four pairs of activation states of the pre and post synaptic neurons by the stimulus: $(V, V)$, $(V, 0)$, $(0, V)$, and (0,0) (where the low spontaneous rate is represented by 0). Note that because we assume a symmetric role for pre and postsynaptic neurons, cases $(V, 0)$ and $(0, V)$ are equivalent, and we consider only the case $(V, 0)$. The number of situations is reduced to six.

- For $J_{ij} = J_0$ and $(\nu_i, \nu_j) = (V, V)$: if the integrated synaptic source (Eq. 4) over the duration of the presentation $t_p$ reaches the potentiation threshold,

  $$(\lambda_+ V^2 - 2\lambda_- V)\left(1 - \exp\left(-\frac{t_p}{\tau_c}\right)\right) > \theta_+$$

  there is a probability $p_+$ of activation of the refresh source, causing a transition of the synaptic value to $J_1$ in the delay period. LTP has occurred. This probability depends on $c_+ = \lambda_+ V^2 - 2\lambda_- V$, $\theta_+$, and the ratio $t_p/\tau_c$.

- For $J_{ij} = J_1$ and $(\nu_i, \nu_j) = (V, 0)$ or $(0, V)$: if

  $$[\lambda_-(V + \nu_s) - \lambda_+ V \nu_s]\left(1 - \exp\left(-\frac{t_p}{\tau_c}\right)\right) > \theta_-$$

  the refresh source will be turned off with probability $p_-$. $J_{ij}$ goes to $J_0$, its low value, in the subsequent delay period. This transition represents LTD. $p_-$ depends on $c_- = \lambda_-(V + \nu_s) - \lambda_+ V \nu_s$, $\theta_-$ and the ratio $t_p/\tau_c$.

- In all other cases no transitions can occur.

Therefore in absence of delay activity, and when the presentation duration is kept fixed, we can represent the synaptic dynamics by a discrete stochastic — a random walk between the two synaptic stable states $J_0$ and $J_1$. This is a familiar situation (Amit and Fusi 1994, Amit and Brunel 1995a), in which uncorrelated stimuli leads to uncorrelated attractors.

## 3.4 Synaptic transitions — Delay activity prior to the presentation

In contrast, when neural activity persists during the delay period, the synaptic dynamics depends on the activation of the pre and post synaptic neurons by the stimulus, but also on the activation of these neurons during the previous delay period. There are now 32 possible situations, depending on whether $J_{ij}$ is above or below threshold before the presentation, and on the pair $(\nu_i, \nu_j)$ during both stimulus presentation and the previous delay period. Since the transient interval during which either old delay and new stimulus-related activities are present is short compared to the presentation interval, the probabilities $p_+$ and $p_-$ will not be much affected by the previous delay activity in the situations described in section 3.3, where LTP or LTD occurs only due to stimulus presentation.

A new LTP transition might occur: if before presentation $J_{ij} = J_0$, and during the transient interval $\tau_E$

$$(\nu_i, \nu_j) = \begin{cases} (v, 0) & \text{during the delay period} \\ (0, V) & \text{during the stimulus presentation,} \end{cases} \tag{6}$$

or

$$(\nu_i, \nu_j) = \begin{cases} (0, v) & \text{during the delay period} \\ (V, 0) & \text{during the stimulus presentation,} \end{cases} \tag{7}$$

and if the integrated source of the synaptic dynamics over $\tau_E$ crosses the potentiation threshold,

$$V(\lambda_+ v - \lambda_-)\left(1 - \exp\left(-\frac{\tau_E}{\tau_c}\right)\right) - \lambda_- v > \theta_+,$$

there is a probability $ap_+$, of activation of the refresh source, which will drive the synaptic efficacy to $J_1$ in the subsequent delay period. $a$ is a function of the ratio $\tau_E/t_p$ and of $v/V$. Typically if the presentation duration is much longer than $\tau_E$ $a \ll 1$.

A similar situation would occur also if $(\nu_i, \nu_j) = (v, v)$ in the delay. However, in this case, the probability of LTP during the previous stimulus presentation is much larger than the one during the short transient period, and can be neglected. The only new situation leading to LTP in presence of delay activity is the one described in (6,7). We will see in the following that this has important consequences for the synaptic matrix in case of significant temporal correlations in the training sequence of stimuli, which in turn will affect significantly the neural dynamics.

To conclude we give a numerical example to illustrate the possible scenarios. We take the background synaptic efficacy $J_0 = 0.04$ mV, $J_1 = 0.15$ mV. The threshold for potentiation is $\theta_+ = 0.04$ mV above $J_0$, and for depression is $\theta_- = 0.04$ mV below $J_1$. The time constants are $\tau_E = \tau_c =$ 10ms, $t_p =$ 200ms. For $\lambda_+ = 5.10^{-4}$ mVs$^2$, $\lambda_- = 4.10^{-3}$ mVs, Fig. 4 shows in the space $(\nu_i, \nu_j)$ the regions where potentiation or depression are possible.

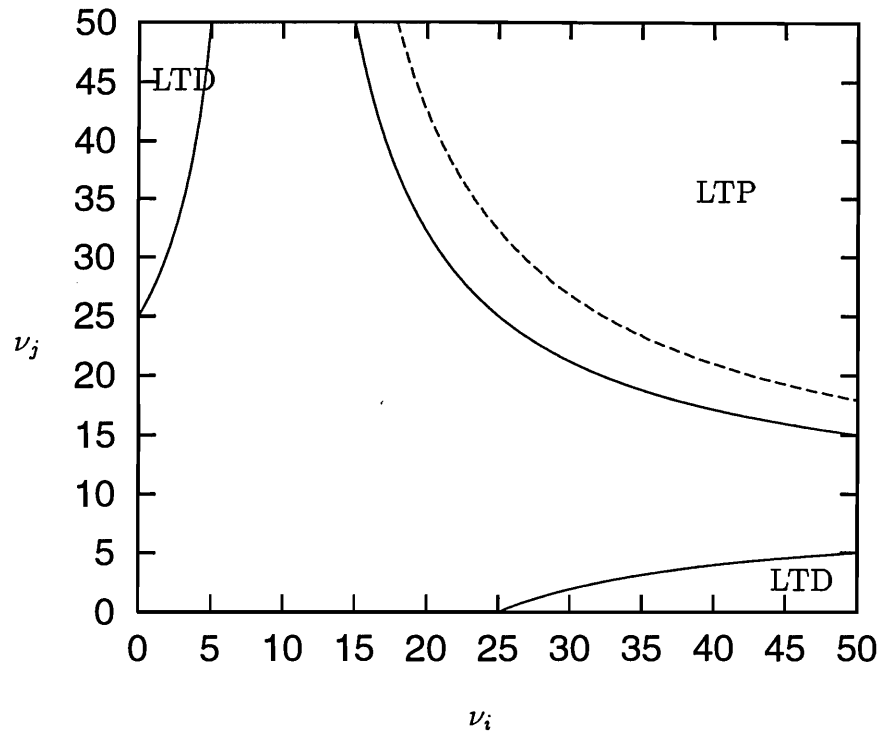


Figure 4: Regions where synaptic transitions occur in the $(\nu_i, \nu_j)$ plane. Frequencies are indicated in spikes per second. Above the dashed line LTP transitions occur due to presynaptic delay activity and postsynaptic activation by the new stimulus. In this case $\nu_i$ is the delay activity prior to presentation of the stimulus.

Three situations leading to possible transitions are schematized in Fig. 5.

To conclude this section we emphasize that one can imagine different scenarios for the occurrence of LTP when one neuron is active during the delay while the other is active during presentation of the next stimulus. For example, it would also naturally occur if the Hebbian source term of the synaptic dynamics described by Eq. (4) depends not on the instantaneous neural activities, but rather on their average over some temporal window. In this section we have argued that in a simple and plausible short term analog dynamics this type of transitions occur naturally. In the following we will not consider anymore the short-term analog synaptic dynamics, but only the resulting stochastic process acting on the two stable synaptic states.

# 4 Training the network with a fixed set of stimuli

We consider the case of a set $E$ of a finite number of stimuli $p$. The initial distribution of excitatory to excitatory synaptic bonds is assumed uniform,

$$\rho_0(J_{ij} = J_1) = g(0), \quad \rho_0(J_{ij} = J_0) = 1 - g(0)$$

for all $(i, j)$. During training the stimuli shown to the network are limited to the set $E$. The learning protocol defines the order in which the stimuli are presented

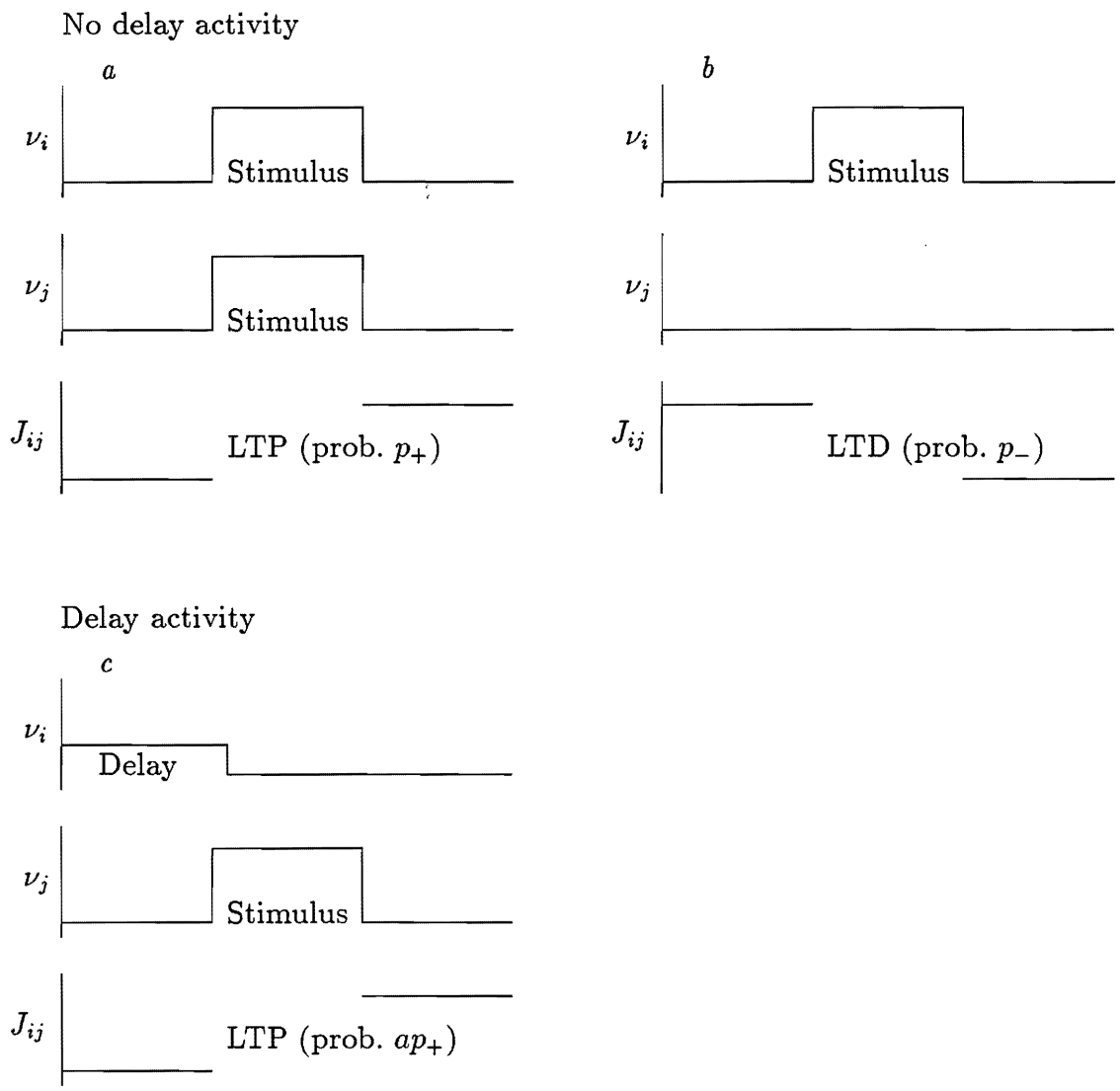


Figure 5: Schematic illustration of synaptic transitions in three situations: time evolution of synaptic efficacy $J_{ij}$ (lower curves), presynaptic activity ($\nu_j$) and postsynaptic activity ($\nu_i$). *a.* Pre and postsynaptic neurons activated by stimulus, synapse initially low. *b.* Presynaptic neuron silent during stimulus, postsynaptic neuron activated, synapse initially high. *c.* Presynaptic neuron activated during stimulus, postsynaptic neuron active in delay, synapse initially low. Note that in all cases one can permute pre and postsynaptic neurons, due to the symmetry of the short-term analog learning dynamics.

to the network. In the following we study the following training protocols:

**A** . Random sequence: at each presentation the stimulus is chosen randomly out of $E$.

**B** . Fixed order: the stimuli are presented in a fixed cyclic order i.e. 1, 2, ..., $p$, 1, and so on. We also study the intermediate situation in which at each time step there is a probability $x$ of showing a randomly chosen stimulus in $E$ instead of the predetermined one. For $x = 1$ one recovers the case of random sequence.

**C** . Random pairs: stimuli in $E$ are organized in $p/2$ pairs. Each stimulus $\mu$ has a paired associate $\bar{\mu}$. The pairs are selected at random. When a pair is chosen both members are shown successively in a random order. We also study the intermediate situation in which at each time step there is a probability $x$ of showing a randomly chosen stimulus instead of one of the paired associates. Again for $x = 1$ the random sequence is recovered.

Protocol B is similar to the protocol of the experiment of Miyashita (1988). In this experiment the *sample* stimuli are shown in a fixed order, while the *match* stimuli are chosen to be the sample with probability 0.5, and a random different stimulus otherwise. Thus it would correspond to protocol B with a probability $x = 0.5$ of showing a random stimulus. Protocol C is similar to the protocol of Sakai and Miyashita (1991). In this experiment the *sample* is a randomly selected stimulus. Then two *match* stimuli are shown: the paired associate and another randomly chosen stimulus.

We consider the case in which the coding level $f$ is very small, so that $fp \ll 1$, but $fC_{EE}$, where $C_{EE}$ is the excitatory to excitatory connectivity, is very large. Consider neurons which are activated by a specific stimulus $\mu$. A fraction $(\exp[-f(p-1)]) \sim 1 - f(p-1)$ of these neurons is not activated by any other stimulus. Thus when $fp \ll 1$, most selective neurons are activated by only one stimulus. We may therefore consider only these neurons, and the network can be functionally divided in $p+1$ sets of neurons. One set corresponds to neurons which are not activated by any stimulus. This set is denoted by $F_0$. The other sets of neurons correspond to neurons which are activated by one of the $p$ stimuli. $F_\mu$ is the population of cells which are activated when stimulus $\mu$ is presented, i.e. $F_\mu = \{i|\eta_i^\mu = 1\}$.

Next we classify accordingly the excitatory-to-excitatory synapses. There are in the network four types of synaptic populations:

- Synapses which connect two neurons activated by the same stimulus. $G_{\mu\mu}$ is the population of all synapses from $F_\mu$ to itself, i.e. $\{(i,j)|\eta_i^\mu = 1, \eta_j^\mu = 1\}$.

- Synapses connecting two neurons activated by two different stimuli. $G_{\mu\nu}$ is the population of synapses from $F_\nu$ to $F_\mu$, i.e. $\{(i,j)|\eta_i^\mu = 1, \eta_j^\nu = 1\}$.

- Synapses connecting a neuron activated by a stimulus to a neuron not activated by any stimulus. $G_{\mu 0}$ is the population of synapses from $F_0$ to $F_\mu$, i.e. $\{(i,j)|\eta_i^\mu = 1, \eta_j^\nu = 0 \text{ for all } \nu.\}$, and $G_{0\mu}$ is the population of synapses from $F_\mu$ to $F_0$, i.e. $\{(i,j)|\eta_i^\nu = 0 \text{ for all } \nu, \eta_j^\mu = 1\}$

- Synapses connecting two neurons none of which is activated by any stimulus. $G_{00}$ is the population of synapses from $F_0$ to $F_0$, i.e. $\{(i,j)|\eta_i^\nu = 0, \eta_j^\nu = 0 \text{ for all } \nu.\}$

To calculate the probability distribution of the synaptic efficacies in each of these populations, as a function of the learning protocol and of the duration of training, we define two units of time: the first corresponds to the interval between two presentations. Time in this unit will be referred to as $t$. The second measure of time $T = pt$, corresponds to the interval between two successive presentations of the same stimulus, for a fixed cyclic sequence as in protocol B. At a given time $t$ $n_\mu(t)$ is the number of times a given stimulus has been presented to the network, while $m_{\mu\nu}(t)$ corresponds to the number of times stimulus $\nu$ has been presented *immediately following the delay activity provoked by stimulus* $\mu$.

The probability distribution of the efficacies in any population $G_{\mu\nu}$ is completely characterized by the probability of the synapse being potentiated, i.e.

$$g_{\mu\nu} = \rho(J_{ij} = J_1|(i,j) \in G_{\mu\nu})$$

since $\rho(J_{ij} = J_0) = 1 - g_{\mu\nu}$ for $(i,j) \in G_{\mu\nu}$. The details of the derivation of these probabilities are given in Appendix.

1. For a synapse in population $G_{\mu\mu}$

$$g_{\mu\mu}(t) = (1-p_+)^{n_\mu(t)} g(0) + 1 - (1-p_+)^{n_\mu(t)}$$

   where $g(0)$ is the initial probability of finding a potentiated synapse. Thus when $n_\mu$, the number of presentations of stimulus $\mu$, becomes large we get $g_{\mu\mu} \to 1$, i.e. all synapses become potentiated.

2. For synapses in population $G_{\mu\nu}$ with $\mu \neq \nu$, the distribution depends not only on $n_\mu$, $n_\nu$ and $n_{\mu\nu}$ but also on when the neighbour presentations were done. There are two simple cases in which the distribution can be calculated. The first is when stimuli $\mu$ and $\nu$ *always* follow each other. In this case the learning protocol can be divided in two intervals: the first corresponds to the absence of delay activity after presentation of a stimulus. After $(n_\mu, n_\nu)$ presentations we have

$$g_{\mu\nu} = (1-p_-)^{n_\mu + n_\nu} g(0),$$

gradually eliminating the potentiated inter-stimulus synapses contained in the initial distribution. In the second interval, delay activity has developed. When $n_{\mu\nu}$ becomes large we obtain (see Appendix for details)

$$g_{\mu\nu} \to \frac{ap_+}{ap_+(1-p_-)+p_-(2-p_-)} \equiv \tilde{a}$$

Another limit case is when $\mu$ and $\nu$ are *never* presented contiguously. In this case the probability of the synapse being potentiated is

$$g_{\mu\nu} = (1-p_-)^{n_\mu+n_\nu} g(0)$$

and therefore vanishes when the number of presentations becomes very large.

In the intermediary situation when joint presentations occur but not systematically we define the relative frequency of the contiguous appearance of the two stimuli

$$\rho_{\mu\nu} = \frac{2n_{\mu\nu}}{n_\mu + n_\nu}$$

The probability of having a potentiated link goes, when the number of presentations becomes very large at fixed $\rho_{\mu\nu}$, to

$$g_{\mu\nu} \to \frac{\rho_{\mu\nu} ap_+}{\rho_{\mu\nu} ap_+(1-p_-)+p_-(2-p_-)} \equiv \tilde{a}(\rho_{\mu\nu})$$

3. For synapses in $G_{0\mu}$ or $G_{\mu 0}$ one has

$$g_{\mu 0} = g_{0\mu} = (1-p_-)^{n_\mu} g(0)$$

and thus the probability of having a potentiated synapse goes to zero in the limit of many presentations of stimulus $\mu$.

4. The last population of synapses is composed of synapses who never see activity in the learning process. These synapses remain unmodified. We will see in the following that these synapses do not play any role in the dynamics of the network.

We are now able to calculate the parameters $g_{\mu\nu}$ for the learning protocols described at the beginning of the section. For each of these learning protocols the probability of occurrence of any stimulus is the same. This probability is $1/p$ where $p$ is the number of stimuli. Thus it is convenient to express the parameters $g_{\mu\nu}$ as a function of $T = pt$. For $G_{\mu\mu}$, $G_{\mu 0}$, $G_{0\mu}$ and $G_{00}$ the distribution is independent of the learning protocol

$$g_{\mu\mu}(T) = (1-p_+)^T g(0) + 1 - (1-p_+)^T$$

$$g_{\mu 0}(T) = g_{0\mu}(T) = (1-p_-)^T g(0)$$

$$g_{00}(T) = g(0)$$

By contrast, the synaptic distributions in populations $G_{\mu\nu}$ for $\mu \neq \nu$ depend rather drastically on the learning protocol. $g_{\mu\nu}$ depends not only on $T$ but also on $\rho_{\mu\nu}$, the frequency of a contiguous presentations of $\mu$ and $\nu$ *connected by a delay activity.* The expression for $g_{\mu\nu}$ is

$$g_{\mu\nu}(T) = (1-p_-)^{T(2-\rho_{\mu\nu})}(1-p_- - ap_+)^{\rho_{\mu\nu}T} g(0) +$$

$$\rho_{\mu\nu} a p_+ \left( \frac{1-(1-p_- - ap_+)^{\rho_{\mu\nu}T}(1-p_-)^{\rho_{\mu\nu}T}}{\rho_{\mu\nu} a p_+ (1-p_-) + p_-(2-p_-)} \right)$$

Recall that the dependence on the learning protocol arises only when persistent delay activity is present in the network.

The next step is to calculate the frequency of contiguous presentation for any pair of stimuli $\rho_{\mu\nu}$, starting from the time at which persistent delay activity became stable in the network. Since during training all stimuli are presented the same average number of times, delay activity appears at the same stage of the learning protocol for all stimuli. We also suppose $p > 2$.

**Protocol A.** (random presentation sequence) For all $\mu \neq \nu$ one has

$$\rho_{\mu\nu} = \frac{2}{p-1}$$

Every pair of stimuli has the same frequency of contiguous occurrence.

**Protocol B.** (fixed presentation sequence) One has

$$\rho_{\mu\mu\pm1} = 1,$$

since $\mu$ and $\mu \pm 1$ always appear contiguously, and

$$\rho_{\mu\nu} = 0 \text{ for all } \nu \neq \mu, \mu \pm 1.$$

Note that in this case, when the number of presentations becomes very large, the synaptic matrix becomes very similar to the matrix used in (Amit et al 1994, Brunel 1994). If there is a probability $x$ of a randomly chosen stimulus between two successive stimuli, we have

$$\rho_{\mu\mu\pm1} = (1-x)^2 + \frac{6x(1-x)}{p} + \frac{2x^2}{p-1}$$

and

$$\rho_{\mu\nu} = \frac{4x(1-x)}{p} + \frac{2x^2}{p-1} \text{ for all } \nu \neq \mu, \mu \pm 1.$$

**Protocol C.** (paired associates) In this case

$$\rho_{\mu\bar{\mu}} = 1,$$

since $\mu$ and $\bar{\mu}$ always occur contiguously.

$$\rho_{\mu\nu} = \frac{1}{p-2}$$

for $\nu \neq \mu, \bar{\mu}$. Again, a paired associate is replaced by a randomly chosen stimulus with probability $x$ we have

$$\rho_{\mu\bar{\mu}} = (1-x)^2 + \frac{6x(1-x)}{p} + \frac{2x^2}{p-1}$$

and

$$\rho_{\mu\nu} = \frac{4x(1-x)}{p} + \frac{2x^2}{p-1} + \frac{(1-x)^2}{p-2} + \frac{2x(1-x)}{p(p-2)} \quad \text{for all } \nu \neq \mu, \bar{\mu}.$$

Thus the different synaptic distributions are now completely determined as a function of the learning stage $T$ and of the learning protocol. They are characterized by the matrix $\rho$ giving the probability of mutual contiguous occurrence of any pair of stimuli in the learning set $E$.

# 5 Learned delay activity distributions

To monitor the neural dynamics we define the average activity of neurons in population $F_\mu$ (neurons driven by stimulus number $\mu$)

$$m_\mu(t) = \frac{1}{fN} \sum_{i \in F_\mu} \nu_i(t) = \frac{1}{fN} \sum_i \nu_i(t) \eta_i^\mu$$

and the average activity of neurons which are not active in response to any stimulus

$$m_0(t) = \frac{1}{(1-fp)N} \sum_i \nu_i(t) \left(1 - \sum_\mu \eta_i^\mu\right)$$

The population-averaged activity in the entire excitatory network is

$$m_E(t) = m_0(t) + f \sum_\mu [m_\mu(t) - m_0(t)]$$

The population-averaged inhibitory activity is

$$m_I(t) = \frac{1}{N_I} \sum_{i \in I} \nu_i^I(t)$$

The average recurrent excitatory current impinging on a neuron of a given population $F_\mu$ (here $\mu$ denotes either a stimulus or 0) is:

$$h_\mu(t) = C\left(J_0 m_E(t) + f(J_1 - J_0)\sum_\nu g_{\mu\nu} m_\nu(t) + (1-fp)(J_1 - J_0) g_{\mu 0} m_0(t)\right) \quad (8)$$

and its variance is

$$\delta_\mu^2(t) = \lambda C\left(J_0^2 m_E(t) + f(J_1 - J_0)^2 \sum_\nu g_{\mu\nu} m_\nu(t) + (1-fp)(J_1 - J_0)^2 g_{\mu 0} m_0(t)\right) \quad (9)$$

The dynamics of the excitatory network is described described by Eqs. (8,9), together with the equations giving the evolution of the means and variances of the depolarizations at the soma of excitatory neurons in populations $F_\mu$. From Eqs. (1,2) it follows that

$$\tau_E \partial_t I_\mu = -I_\mu + I_\mu^{ext} + h_\mu - C_{EI} J_{EI} m_I, \quad (10)$$

and

$$\frac{\tau_E}{2}\partial_t\left(\sigma_\mu^2\right) = -\sigma_\mu^2 + (\sigma_\mu^{ext})^2 + \delta_\mu^2 + C_{EI} J_{EI}^2 m_I. \quad (11)$$

The terms appearing on the right hand side of Eqs. (10,11) are: the decay term; the external contribution; the recurrent excitatory contribution, given by Eqs. (8,9); and the inhibitory contribution.

The corresponding equations for the inhibitory neurons are given by

$$\tau_I \partial_t I_I = -I_I + I_I^{ext} + C_{IE} J_{IE} m_E - C_{II} J_{II} m_I, \quad (12)$$

and

$$\frac{\tau_I}{2}\partial_t\left(\sigma_I^2\right) = -\sigma_I^2 + (\sigma_I^{ext})^2 + C_{IE} J_{IE}^2 m_E - C_{II} J_{II}^2 m_I \quad (13)$$

In Eqs. (12,13), the terms appearing on the right hand side are again: the decay term; the external contribution; the recurrent excitatory contribution; and the inhibitory contribution. The average activity in each population is in turn given by

$$m_\mu = \phi_E(I_\mu, \sigma_\mu), \quad (14)$$

and

$$m_I = \phi_I(I_I, \sigma_I), \quad (15)$$

where the transduction functions $\phi_\alpha$ ($\alpha = E, I$) are given by Eq. 3.

To obtain the delay activity after presentation of a given stimulus $\mu$ at learning stage $T$ we proceed as follows:

1. Initially all neurons have their stable spontaneous activity. Only background external currents are present.

2. Stimulus number $\mu$ is presented by injecting into neurons of population $\mu$ a 'selective' external current above the background one. Neurons in this population are driven by the selective currents well above their spontaneous rates. Presentation lasts 100ms ($= 10\tau_E$).

3. At the end of the presentation the 'selective' external currents are removed and only background external afferents remain. After a short transient all neurons reach a steady-state delay activity, which persists indefinitely.

We choose the following parameters: the synaptic transition probabilities are: $p_+ = p_- = 0.2$, the neural parameters are as in section 2. The background synaptic efficacy is $J_0$ =0.04mV, while the potentiated synaptic efficacy is $J_1 = 0.15$mV. The synaptic transition probability in the case of contiguous delay activity and stimulus activation $ap_+$, is given by the following values of $a$: $a = 0.02$ and $a = 0.05$. We use $p = 50$ stimuli, each stimulus activating a fraction $f = 0.01$ of the excitatory neurons in the network (Brunel 1994). We have not explored the parameter space. Instead we have chosen a particular set of parameters to exhibit a case of good agreement with the experimentally observed delay activities in IT cortex of performing monkeys.

## 5.1 Protocol A

Stimuli are shown in a random sequence. The upper part of Fig. 6 shows the evolution of delay activities as a function of the learning stage (number of presentations per stimulus) for neurons in the population corresponding to the stimulus presented (diamonds), and neurons in populations corresponding to other stimuli (crosses). It shows that there is a critical learning stage $T_c$, here $T_c = 11$, (minimal number of presentations per stimulus for the creation of an attractor) beyond which selective delay activity appears. This critical learning stage is similar to the critical synaptic parameter of Amit and Brunel (1995b). Before $T_c$, neurons which are active during the presentation of any stimulus see their spontaneous activity slightly increase with $T$. This spontaneous activity is of order 3-4 $s^{-1}$. After $T_c$ the neurons representing the shown stimulus have an elevated delay activity of the order of 20-35 $s^{-1}$. Other excitatory neurons remain at spontaneous activity levels. The critical stage $T_c$ depends on the learning speed, which is controlled by the probabilities $p_+$ and $p_-$. The lower part of figure 6 shows the corresponding evolution of the activity of inhibitory neurons (crosses), which also slightly increases with learning, and of other excitatory neurons not activated in any stimulus (diamonds), which decreases from 3 to 2 $s^{-1}$. In this case delay activities are uncorrelated since they simply reflect the structure of uncorrelated stimuli.

## 5.2 Delay activities for protocol B

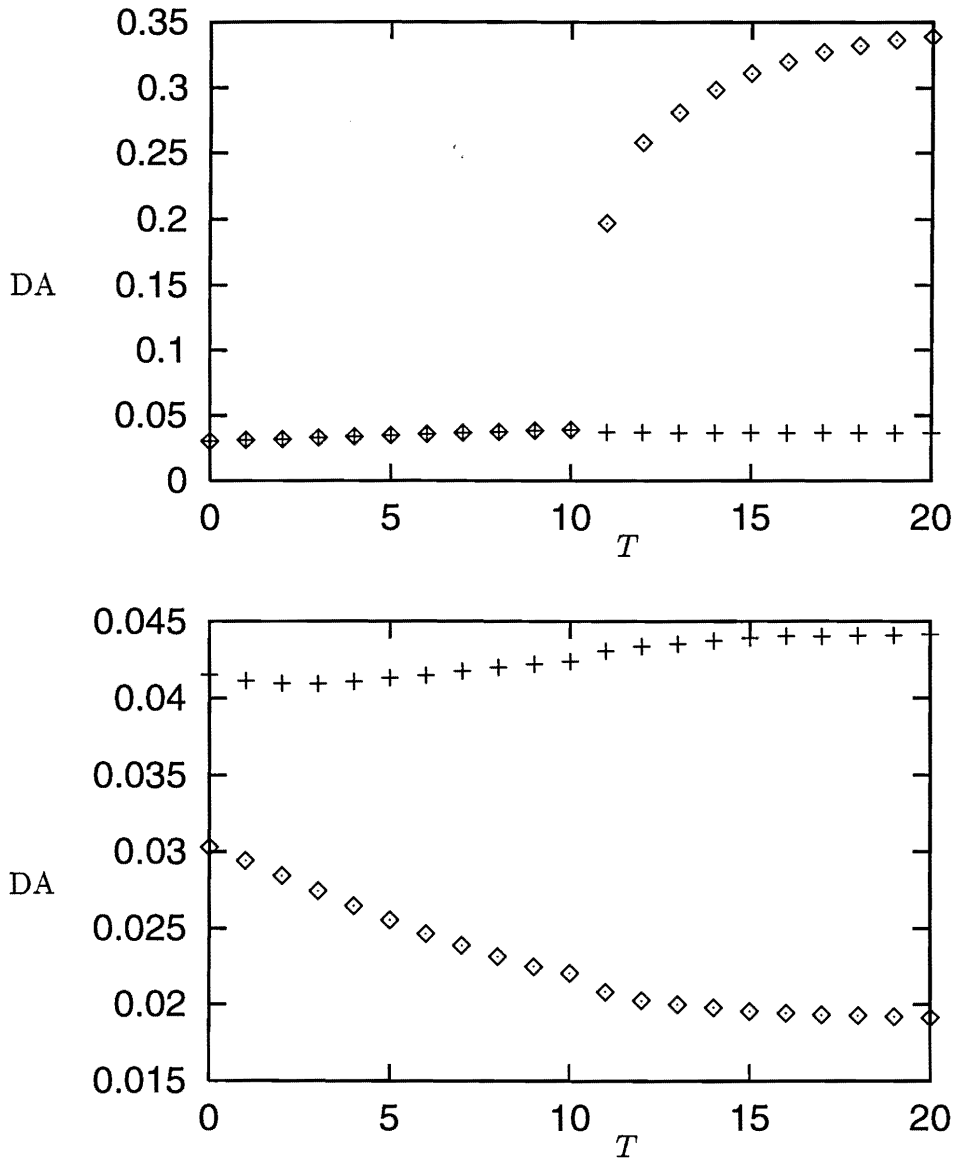


Figure 6: Upper figure: delay activity (DA) of neurons coding for the shown stimulus ($\diamond$) and of neurons coding for other stimuli (+), as a function of the learning stage $T$. Lower figure: delay activity of inhibitory neurons (+) and other excitatory neurons ($\diamond$). Activity is in units of $1/\tau_E$, i.e. 100 $s^{-1}$.

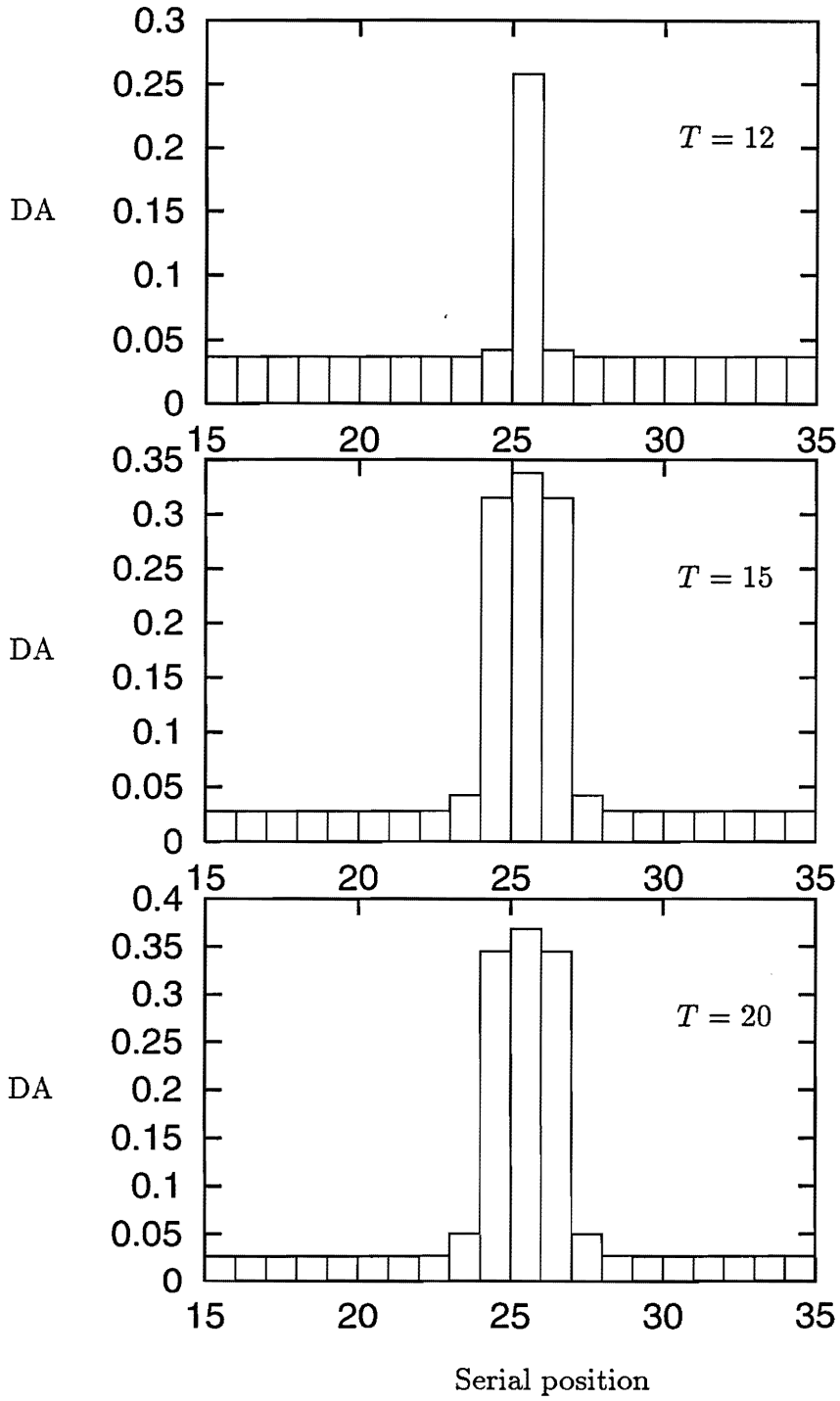


Figure 7: Delay activity of a cell in population $F_{25}$, as a function the serial position of the shown stimulus, for $a = 0.05$ and three values of the learning stage $T$, indicated in the figure. The cell is active in the delay following stimulus 25 but also in the delays following the presentation of its neighbors. These figures can be compared with Fig. 3a of (Miyashita 1988).

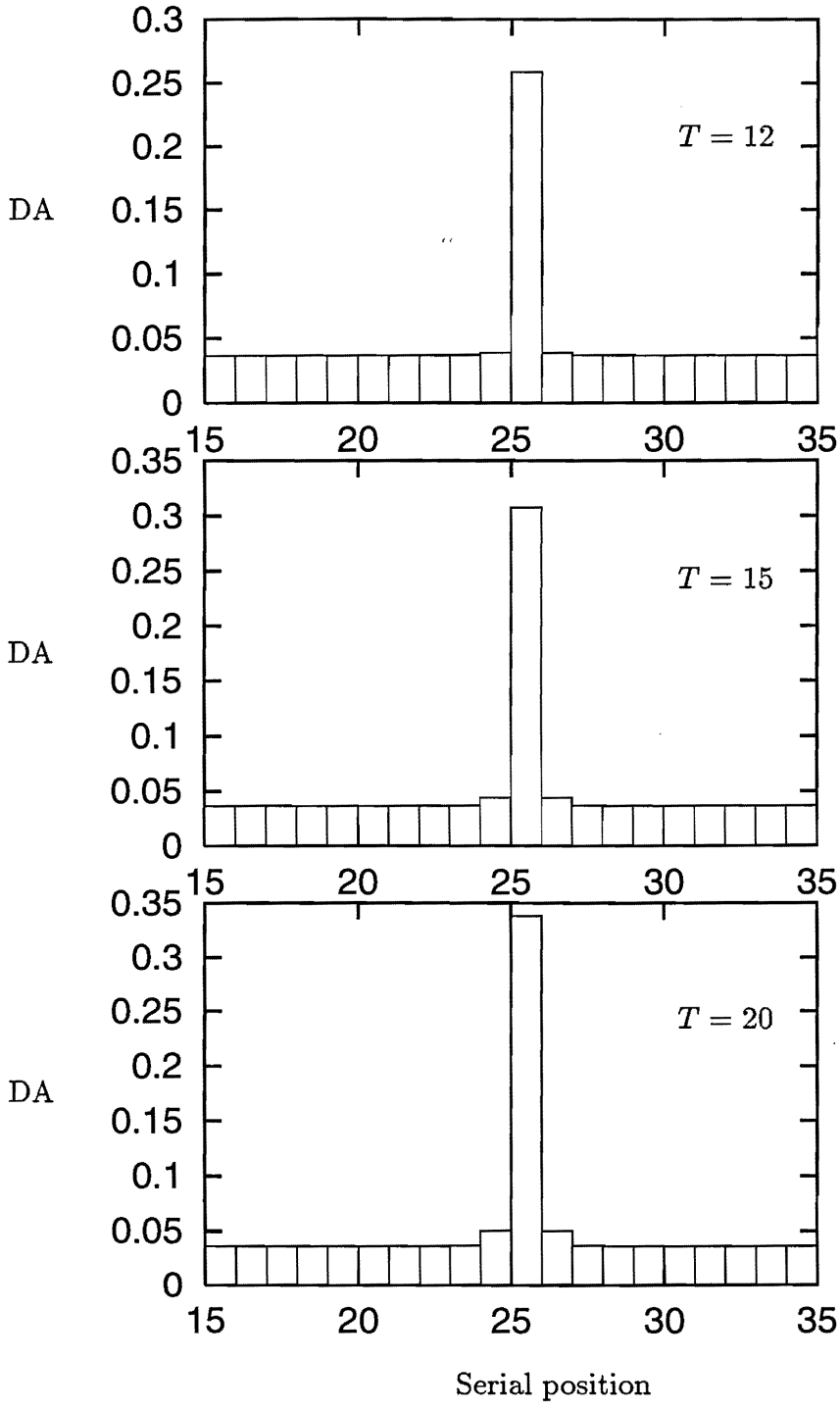


Figure 8: Same as figure 7, but with $a = 0.02$.

Stimuli are presented in a fixed order. Before $T_c$, since there is no delay activity in the system, the neural rates are independent of the order of presentation. Immediately after $T_c$, uncorrelated attractors develop as in the case of protocol A. Presentation of a given stimulus $\mu$ activates neurons of the corresponding population, and this activity is maintained after removal of the stimulus, because synapses connecting these neurons have been sufficiently potentiated. After a while, activity in these neurons also provokes an increase in the activity in neurons in the populations of the neighbouring stimuli, i.e. $\mu + 1$ and $\mu - 1$, since synapses connecting these populations to $F_\mu$, i.e. synapses of $G_{\mu\mu\pm1}$ have now an increased average efficacy. This activity can then propagate to further neighbours, i.e. $\mu \pm 2$, and so on. However, the inhibition controls the overall level of activity in the excitatory network and therefore the activation spreads only to a few neighbours. This activation is also controlled by the parameter $a$, which characterizes the magnitude of the strength of synapses of $G_{\mu\mu+1}$ relative to those of $G_{\mu\mu}$. Depending on this parameter $a$, there exist two regimes, one of low correlation, the other of high correlation.

- High correlation (Fig. 7, $a = 0.05$): after $T = 15$ learning cycles the activation of a neuron coding for a given stimulus in the delay following the presentation of its neighbours becomes of the order of its activation of in the delay following the stimulus itself. When learning proceeds more neighbours see their neurons increase significantly their delay activity. In this case the correlations between two attractors corresponding to neighbor stimuli are very high.

- Low correlation (Fig. 8, $a = 0.02$): the activity of neurons in neighbouring populations, though increased with respect to the other populations, remain low compared to the activity of neurons that represent the shown stimulus. Correlations between two representations of neighbour stimuli remain relatively weak.

In absence of stable spontaneous activity, (as was the case in Brunel 1994) the structure of the delay activity is always as in Fig. 7 (highly correlated delay activities). The presence of a stable spontaneous activity allows for reverberations in which neurons coding for stimuli which are neighbours of the presented stimulus remain at low levels of activity (compared with the activation of neurons coding for the presented stimulus), though it is significantly higher than their spontaneous activity.

It is easy to calculate correlations as well as rank correlation coefficients between the delay activities provoked by different stimuli (see Brunel 1994). Qualitatively these correlations are a decreasing function of the distance in the serial position of the stimuli that provoked the delay activities. These correlations decay to zero (or to negative values in the case of rank correlations) at a distance corresponding to

the number of populations of cells activated above spontaneous levels in a given attractor. For example, in Fig. 7 the correlations would be significant up to a distance of 5 in serial position.

### 5.3 Protocol C - paired associates

In the case of paired associates the situation is qualitatively similar to protocol B, except for the fact that now only neurons coding for the shown stimulus and its paired associate are activated in the delay period. Also in this case we can identify two regimes, with strong or weak correlation between delay activities corresponding to the pair associates. The main difference is that now, in the strongly correlated regime, the delay activity of paired associate neurons is equal to the delay activity of the neurons coding for the shown stimulus. Therefore the network has formed attractors which do not correspond anymore to the individual pictures, but rather to the pairs of pictures. This can be seen in Fig. 9 ($a = 0.05$) at learning stage $T = 15$. By contrast in Fig. 10 the representations of paired associates become correlated with learning, but remain distinct. Note the similarity of this figure with one of the cells shown in (Sakai and Miyashita 1991). However the comparison is not direct. Sakai and Miyashita (1991) gives the activity of cells during presentation of the stimulus. The corresponding delay activity distributions, presented here, are not reported. The analysis predicts that the delay activity provoked by two paired associates should be significantly correlated or even become equal. Note that the formation of similar pair-coding attractors has also been observed in a model with a fixed synaptic matrix (Parga 1994).

## 6 Discussion

In this paper we have discussed an explicit, plausible learning process in a recurrent neural network, which in the presence of delay activity, implements the memory of the context of the learned stimuli in the synaptic matrix. In the case of stimuli shown in a fixed sequence during training, this synaptic matrix is found to be qualitatively similar to the matrix that was used in (Amit et al 1994, Brunel 1994). With such a learning process it is possible to determine the statistical properties of the synaptic matrix as a function of the learning stage and the learning protocol. With the network composed of excitatory and inhibitory cells described in (Amit and Brunel 1995b), whose stable state in absence of learning is a state in which neurons have a spontaneous activity of the order of 1 spike per second, it is in turn possible to determine the statistical properties of the delay activities, again as a function of the learning stage and the learning protocol. In the only case in which to our knowledge experimental data is available (Miyashita 1991) we recover the results of (Amit et al 1994, Brunel 1994) which are in good agreement with the experiment. Furthermore the analysis allows to predict either the evolution of the

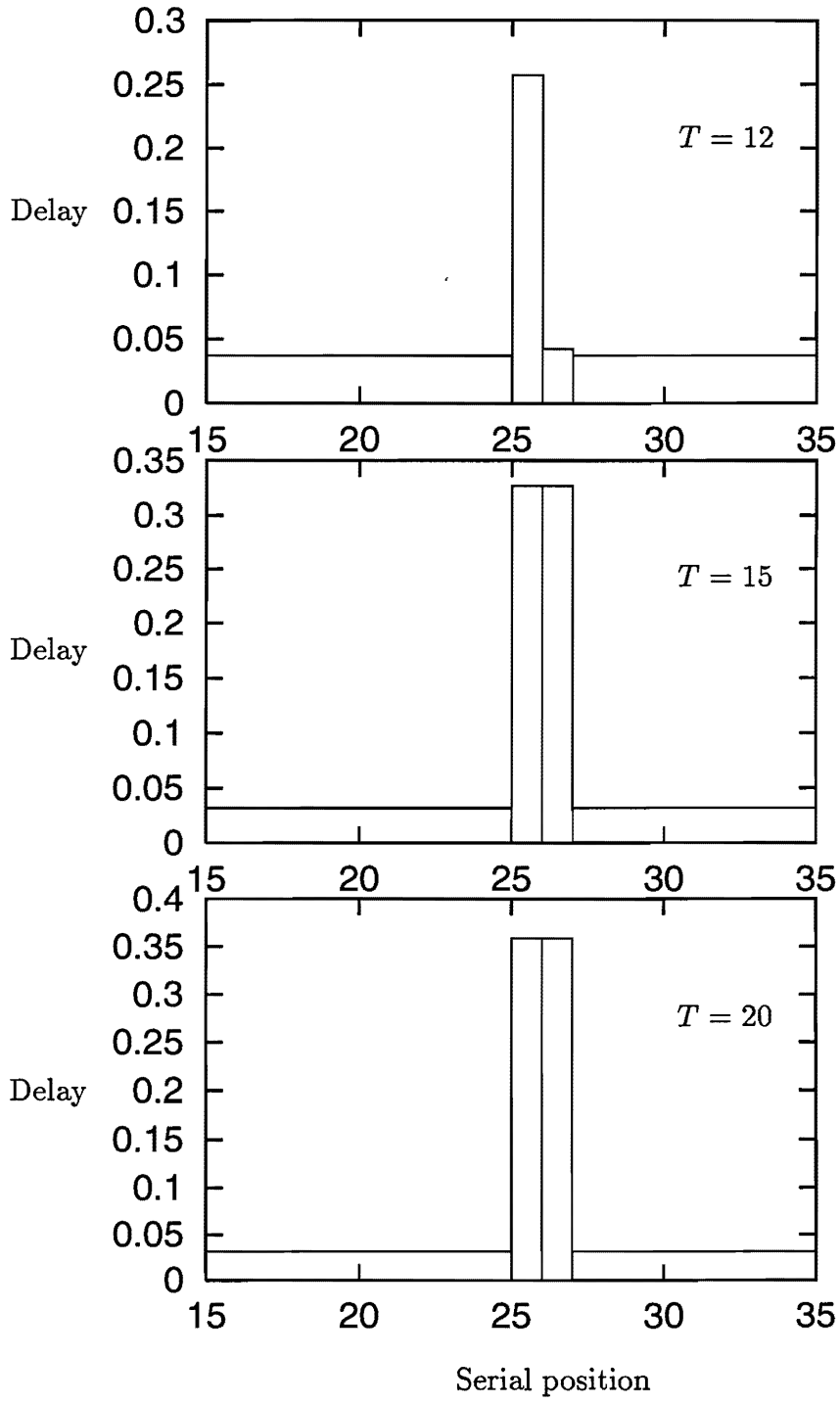


Figure 9: High Delay activity of a cell in population $F_{25}$, as a function of the serial position of the shown stimulus, for $a = 0.05$. The cell is active in the delay following stimulus 25 but also after its paired associate (stimulus 26) is presented.

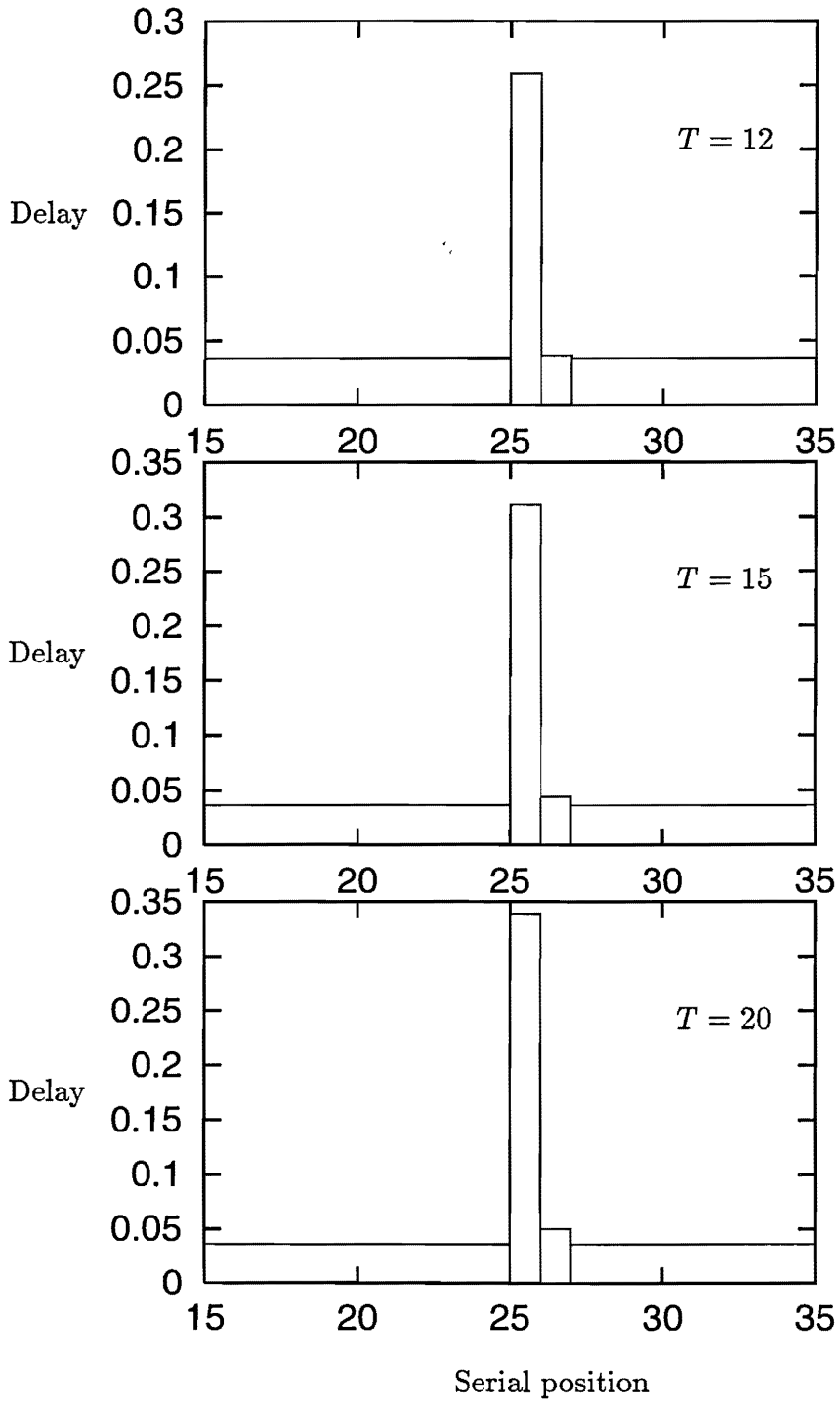


Figure 10: Same as fig. 9, but for $a = 0.02$.

correlations during learning or the dependence of the correlations with the learning protocol.

There are a number of tests of the theory that can in principle be done with visual memory experiments.

1. The time of occurrence of selective delay activity should not depend on the learning protocol, i.e. on the way stimuli are presented.

2. Delay activities corresponding to uncorrelated stimuli should initially be uncorrelated.

3. Correlations between delay activities should only depend on the order of presentation *after* the appearance of selective delay activity in the network, and not on the order of presentation prior to delay activity. For example if stimuli are shown in a fixed order before the appearance of selective delay activity, but in a random order afterwards, the attractors should be uncorrelated.

We turn now to a brief discussion of the elements of the model. Excitatory and inhibitory cells are integrate-and-fire neurons described by the statistics of their input currents and their output firing frequency (Amit and Brunel 1995b). We emphasize that this model, roughly accounts for the average spontaneous and selective activities observed in the visual memory experiments. Last, though the average delay activities themselves do depend on the details of the model neuron, the *correlations* between the attractors of the system seem largely independent on the details of the single neuron. Large-scale simulations of networks of integrate-and-fire neurons are currently under way to confirm that these correlations are preserved if one considers networks of spiking neurons rather than neurons described by firing rates.

The implementation of temporal correlations between stimuli in the synaptic matrix depends crucially on a mechanism leading to long term potentiation when delay activity in one neuron connected by a synapse is immediately followed by stimulus-provoked activity in the other neuron connected by that synapse. This simple mechanism leads to the implementation of such correlations. In this paper this mechanism - and the whole synaptic process - was supposed to be symmetric in pre and post synaptic neurons. This assumption of symmetry was taken for simplicity, but it is not necessary. In fact experimental data suggests LTP can be induced when postsynaptic activity follows presynaptic activity by 100ms (Levy and Steward 1983, Gustafsson et al 1987), but on the other hand, if postsynaptic activity precedes presynaptic activity, LTP does not occur. The formalism developed in this paper can easily be generalized to such an asymmetric situation. This issue will be considered in a future work.

## Acknowledgements

I am grateful to Daniel Amit and Stefano Fusi for many discussions, and to Daniel Amit and Paolo del Giudice for the many detailed comments on a previous version of this manuscript. This work was supported by a fellowship of Programme Cognisciences, CNRS, France.

## Appendix A. Synaptic distributions

1. For a synapse in population $G_{\mu\mu}$: at each presentation of stimulus $\mu$, a synapse which is in its low state, has a probability $p_+$ of making a transition to the potentiated state. Thus after $n_\mu(t)$ presentations

$$g_{\mu\mu}(t) = (1-p_+)^{n_\mu(t)} g(0) + 1 - (1-p_+)^{n_\mu(t)}$$

where $g(0)$ is the initial probability of finding a potentiated synapse.

2. For synapses in population $G_{\mu\nu}$ with $\mu \neq \nu$, the situation is somewhat more complicated, since the distribution depends not only on $n_\mu$, $n_\nu$ and $n_{\mu\nu}$ but also on when the neighbour presentations were done. There are two simple cases in which the distribution can be calculated. The first is when stimuli $\mu$ and $\nu$ *always* follow each other. In this case the learning protocol can be divided in two intervals: the first corresponds to the absence of delay activity after presentation of a stimulus. At each presentation of stimuli $\mu$ or $\nu$, potentiated synapses have a probability $p_-$ of making a transition to the low state. Thus after $(n_\mu, n_\nu)$ presentations we have

$$g_{\mu\nu} = (1-p_-)^{n_\mu+n_\nu} g(0),$$

In the second interval, delay activity has developed. When a contiguous presentation of $\mu$ and $\nu$ occurs there is a probability $ap_+$ for low synapses of making a transition to the high state. Thus after $n_{\mu\nu}$ occurrences of the contiguous presentation of stimuli $\mu$ and $\nu$ separated by the delay period we have

$$g_{\mu\nu} = (1-p_-)^{n_\mu+n_\nu-n_{\mu\nu}}(1-p_- - ap_+)^{n_{\mu\nu}} g(0) + ap_+ \left( \frac{1-(1-p_- - ap_+)^{n_{\mu\nu}}(1-p_-)^{n_{\mu\nu}}}{ap_+(1-p_-) + p_-(2-p_-)} \right)$$

When $n_{\mu\nu}$ becomes large we have

$$g_{\mu\nu} \to \frac{ap_+}{ap_+(1-p_-) + p_-(2-p_-)} \equiv \tilde{a}$$

Another limit case is when $\mu$ and $\nu$ are *never* presented contiguously. In this case the probability of the synapse being potentiated is

$$g_{\mu\nu} = (1-p_-)^{n_\mu+n_\nu} g(0)$$

and therefore vanishes when the number of presentations increases.

In the intermediary situation when joint presentations occur but not systematically we use an interpolation in the relative frequency of the contiguous appearance of the two stimuli

$$\rho_{\mu\nu} = \frac{2n_{\mu\nu}}{n_\mu + n_\nu}$$

This expression is

$$g_{\mu\nu} = (1-p_-)^{n_\mu+n_\nu-n_{\mu\nu}}(1-p_- - ap_+)^{n_{\mu\nu}} g(0) +$$

$$\rho_{\mu\nu} a p_+ \left( \frac{1-(1-p_- - ap_+)^{n_{\mu\nu}}(1-p_-)^{n_{\mu\nu}}}{\rho_{\mu\nu} a p_+ (1-p_-) + p_-(2-p_-)} \right)$$

and interpolates between the two preceding limit cases. The probability of having a potentiated link goes, when the number of presentations becomes very large at fixed $\rho_{\mu\nu}$, to

$$g_{\mu\nu} \to \frac{\rho_{\mu\nu} a p_+}{\rho_{\mu\nu} a p_+ (1-p_-) + p_-(2-p_-)} \equiv \tilde{a}(\rho_{\mu\nu})$$

3. For synapses in $G_{0\mu}$ and $G_{\mu 0}$, presentation of stimulus $\mu$ causes depression with probability $p_-$, and after $n_\mu$ presentations one has

$$g_{\mu 0} = g_{0\mu} = (1-p_-)^{n_\mu} g(0)$$

and thus the probability of having a potentiated synapse goes to zero in the limit of many presentations of stimulus $\mu$.

# References


Amit DJ 1995 The Hebbian paradigm reintegrated: Local reverberations as internal representations, *BBS*, to be published

Amit DJ and Brunel N 1995a Learning internal representations in an attractor neural network, *Network*, in press.

Amit DJ and Brunel N 1995b Global spontaneous activity and local structured (learned) delay activity in cortex, submitted.

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

Amit DJ and Fusi S 1994 Dynamic learning in neural networks with material sysnapses, *Neural Computation*, **6**

Amit DJ and Tsodyks MV 1991 Quantitative study of attractor neural network retrieving at low spike rates I: Substrate – spikes, rates and neuronal gain *Network* **2** 259

Atick JJ 1992 Could information theory provide an ecological theory of sensory processing? *Network* 3, 213

Badoni D, Bertazzoni S, Buglioni S, Salina G, Amit DJ and Fusi S 1995, Electronic implementation of an analog attractor neural network with stochastic learning, *Network*, in press.

Barlow HB 1961 Possible principles underlying the transformation of sensory messages, in *Sensory Communication*, Rosenblith W.A. (ed), MIT press

Braitenberg V and Schuz A 1991 *Anatomy of the cortex*, (Springer-Verlag, Berlin)

Brunel N 1994 Dynamics of an attractor neural network converting temporal into spatial correlations, *Network*, **5** 449

Brunel N and Fusi S 1995, in preparation.

Dehaene S, Changeux JP 1989 A simple model of prefrontal cortex function in delayed response tasks *J. Cognit. Neurosci.* **1** 3

Fuster JM 1973 Behavioural electrophysiology of the prefrontal cortex, *J. Neurophysiol.* **36** 61

Fuster JM 1995 *Memory in the Cerebral Cortex*, (MIT Press, Cambridge)

Fuster JM and Jervey JM 1981 Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli, *Science* **212** 952

Goldman-Rakic PS 1987 Circuitry of primate prefrontal cortex and regulation of behaviour by representational knowledge. In *Handbook of Physiology*, Vol. 5, 373 (Bethesda,MD: American Physiological Society)

Griniasty M, Tsodyks MV and Amit DJ 1992 Conversion of temporal correlations between stimuli to Spatial correlations between attractors, *Neural Computation* **5** 1

Gustafsson B, Wigstrom H, Abraham WC and Huang YY 1987 Long-term potentiation in the hippocampus using depolarizing current pulses as the conditioning stimulus to single volley synaptic potentials, *J. Neurosci.* **7**, 774

Komatsu Y, Nakajima S, Toyama K, Fetz E 1988 Intracortical connectivity revealed by spike-triggered averaging in slice preparations of cat visual cortex, *Brain Res.* **442** 359

Levy WB and Steward D 1983 Temporal contiguity requirements for long-term associative potentiation or depression in the hippocampus, *Neurosci.* **8**, 791

Linsker R 1989 An application of the principle of maximum information preservation to linear systems, in *Advances in Neural Information Processing Systems 1*, Touretzky DS (ed), Morgan-Kauffman

Mason A, Nicoll A, Stratford K 1991 Synaptic transmission between individual pyramidal neurons of the rat visual cortex *in vitro*, *J. Neurosci.*, **11** 72

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

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

Niki H 1974 Prefrontal unit activity during delay alternation in the monkey, *Brain Res.* **68** 185

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

Tanaka K 1992 Inferotemporal cortex and higher visual function, *Current Biology*

Parga N Private communication.

Ricciardi LM 1977 *Diffusion processes and Related topics on biology* (Springer-Verlag, Berlin)

Willshaw D Buneman O P and Longuet-Higgins H 1969 Non-holographic associative memory, *Nature*, **222** 960

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

Zipser D, Kehoe B, Littlewort G and Fuster J 1993 A spiking network model of short-term active memory, *J. Neurosci.* **13** 3406