# Dynamics of an attractor neural network converting temporal into spatial correlations

**Nicolas Brunel**
**INFN, Sezione di Roma, Istituto di Fisica**
**Universita di Roma, La Sapienza, Ple Aldo Moro, Roma**

**Abstract**

The dynamics of a model attractor neural network, dominated by collateral feedback, composed of excitatory and inhibitory neurons described by afferent currents and spike rates, is studied analytically. The network stores stimuli learned in a temporal sequence. The statistical properties of the delay activities are investigated analytically under the approximation that no neuron is activated by more than one of the learned stimuli, and that inhibitory reaction is instantaneous. The analytic results reproduce the details of simulations of the model in which the stored memories are uncorrelated, and neurons can be shared, with low probability, by different stimuli. As such, the approximate analytic results account for delayed match to sample experiments of Miyashita in the inferotemporal cortex of monkeys. If the stimuli used in the experiment are uncorrelated, the analysis deduces the mean coding level $f$ in a stimulus (i.e. the mean fraction of neurons activated by a given stimulus) from the fraction of selective neurons which have a high correlation coefficient, of $f \sim 0.0125$. It also predicts the structure of the distribution of the correlation coefficients among neurons.

# 1 Introduction

The experiments of the Miyashita group [1, 2, 3] of single unit recordings in the inferotemporal cortex of monkeys trained to perform delayed matching to sample tasks have provided evidence for attractor dynamics in a module of the region under observation. Similar phenomena have been observed in pre-frontal cortex (see e.g. Ref. [4]). In one of the experiments [2] significant correlations in the internal representations of stimuli, chosen to be uncorrelated, have been obtained when those had been presented during training in a fixed sequence. Given the potential of these experiments of bridging the gap between experiments and theory, several studies have been made in order to obtain these correlations in a model attractor neural network. A detailed theoretical model for such experiments can much enhance their implications, as well as provide precise clues for tests and extensions.

A first step in this direction, stimulated by those experiments, was made in a recent study [5]. It was found by simulations that a simple attractor neural network which connects, in its synaptic structure, information about contiguous stimuli learned in a sequence, has correlated delay activities even though the learned stimuli are uncorrelated. In this model network the delay activity distribution (an attractor, or reverberation) provoked in the neural assembly by the presentation of one of the uncorrelated stimuli, is correlated with the delay activity corresponding to other stimuli until a separation of several patterns in the sequence of the learned patterns, although the synaptic matrix connects only information about consecutive (nearest neighbors) in the sequence. The correlation distance of the attractors, as well as the amplitudes of the correlations are robust to the parameters of the model. 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 among neurons) correlations of the internal representations of the different stimuli. In other words, this is an embryo of context sensitivity (see e.g. Ref. [6]). However this model was too simple to obtain a quantitative agreement with the experiment, since neurons were discrete.

The second step was to model the phenomenon in a network of more realistic elements [7]. The idea was to inquire into the domain of validity and robustness of the surprising result found in [5], and, on the other hand, to investigate the characteristics that would bring the model to the level of quantitative agreement with as much information as that given by the short accounts of the experiments.

The model simulated in Ref. [7] consists of a network of integrate-and-fire neurons operating in the presence of high levels of non-selective uncorrelated noise due to spontaneous activity (see e.g. Ref. [8]). The neuron is represented by its current to spike rate transduction function, which includes the effect of noise due to spontaneous activity [8]. 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 expressed. 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, like in Ref. [5]. Inhibition is taken to have fixed synapses and its role is to react rapidly in proportion to the mean level of activity in the excitatory network. In this way it controls the overall activity in the network.

The delay activities are investigated by presenting to the neural module (cell assembly) one of the uncorrelated stimuli as a set of afferent currents into a subset of the 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 collateral part of the synaptic matrix. 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. In [7] it was reported that the phenomenon of the conversion of temporal correlations (contiguity of stimuli in the training sequence) into spatial correlations of neural delay activity distributions in a cell assembly, persists when the model network is composed of quasi-realistic neural elements. Furthermore this new model is in quantitative agreement with all the available experimental data [7, 2], and makes several experimental predictions.

Before we introduce the present work we briefly summarize the experiments of Ref. [2], since we adress it in detail in the following.

## Summary of the experiments of Ref. [2]

In the experiment described in Ref. [2] two monkeys are trained to perform a visual memory task, in which they have to memorize a sample stimulus during a 16 seconds delay period and then to decide whether a second stimulus is the same as the sample. Two sets of 97 computer generated fractal colour patterns are used. The first set is repeatedly used during the training session. Patterns in this set are referred to as learned stimuli, while those from the other set are referred to as new stimuli. Both sets are used during the recording session.

Miyashita finds that a few of the 97 learned stimuli reproducibly activate a particular neuron with high-frequency sustained discharges during the delay period of the task. The optimal set of stimuli varies from cell to cell. By contrast, the 97 new patterns produce only weak delay responses. Of the 206 neurons recorded in the $TE_{av}$ area of IT cortex of the over-trained monkeys, 57 neurons exhibit these selective delay discharges. In the following such cells are referred to as selective cells.

Then Miyashita investigates the effect of a fixed order presentation of the patterns during the training session according to an arbitrarily attached serial position number (SPN). He finds that the effective responses to the learned stimuli cluster along the SPN (see Fig. 3b of Ref. [2]), and that the clustering is not due to an artefact in the testing procedure because the responses simultaneously obtained from the new stimuli are not clustered. In the 57 selective cells, the responses to the learned stimuli are significantly correlated in the neighbours of the SPNs. The nearest-neighbour correlations for the learned stimuli differ from cell to cell, and are significant in 28 of the 57 cells according to Kendall's correlation test, whereas those for the new stimuli are not significant in any cell. In the following these cells with significant nearest-neighbour coefficient are referred to as cells with a high correlation coefficient. The measured correlations are Kendall rank correlation coefficients (see Sec. 4.2 for their definition). They are shown in Fig. 3c of Ref. [2]. The coefficients along the SPNs in these 28 cells differs significantly up to the 6th neighbour from that of the new stimuli. Miyashita concludes that cells in this area can associate pictorial patterns for 6 consecutive neighbours during learning.

## The present work

In the following we present an analytical investigation of the dynamics of the network described in [7] and of the statistical properties of the delay activities in simplified conditions. The main simplifications made through the paper are

- Each excitatory neuron is activated by at most one of the stimuli presented to the network. This is a good approximation only in the limit of a very low coding rate, i.e. when the fraction of neurons driven by each stimulus is low.
- Inhibition is taken to be instantaneous, i.e. the inhibitory currents afferent on any excitatory neuron at any time depend on the instantaneous mean activity of the excitatory network at the same time.
- In sect. 3.2 we will study the case in which all gain functions are taken to be threshold linear [9]. This makes possible the explicit calculation of the delay activities.

The synaptic matrix used is of a Willshaw type [10], i.e. synapses take only one of a few possible values. We have also studied the case in which such a matrix is generated dynamically in a process of learning. It will be described elsewhere [11].

The results of the analysis agree with the simulations in which the above simplifications did not hold. They can be directly confronted with the experimental data of Ref. [2]. If the stimuli used in the experiment are uncorrelated, the analysis produces a relation between the number of selective neurons and the mean number of neurons with a high correlation coefficient. This in turn produces a prediction of a coding level (i.e. the mean fraction of neurons activated by a given stimulus) $f \sim 0.0125$. Furthermore, the analysis provides not only the mean correlations of reverberations over all neurons in a sample, but also the distribution of the correlation coefficients over the neurons. This is yet another prediction.

The article is organized as follows: In section 2 we describe in more detail the model network and its elements. In section 3 we solve the dynamics, first in a toy model in which all transduction functions are threshold-linear, then with more realistic current-to-rate transduction functions. In section 4 the statistical properties of the attractors are presented and are confronted with the results of the experiments in inferotemporal cortex[2]. In the last section we discuss our results.

# 2 The model network

## 2.1 The excitatory network

The network is composed of $N$ excitatory neurons and an associated inhibitory network. An excitatory neuron $i$ ($i = 1, \ldots, N$) is characterized at time $t$ by its incoming current $I_i(t)$ and its firing rate $V_i(t)$. Its afferent current is composed of feed-back (collateral) from the other excitatory neurons; of hyper-polarizing current from the inhibitory network $-T_i(t)$ and an eventual external current $H_i^{ext}(t)$ representing the stimulus. The dynamics of the excitatory neurons is

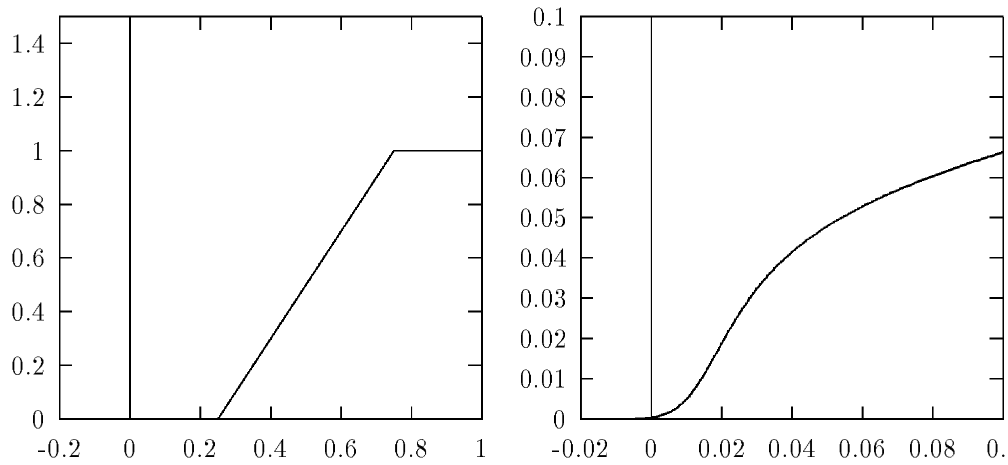


Figure 1: Excitatory transfer functions. Left: threshold-linear, with $\theta = 0.25$ and $g = 2$. Right: 'integrate-and-fire', with $\theta = 2.04$, $\tau = 4$, $\mu = 2$ and $\sigma = 0.015$.

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

The incoming current into neuron $i$ is converted into a spike rate via

$$V_i = \phi_{exc}(I_i), \tag{2}$$

where $\phi_{exc}$ is the current-to-rate transduction function for an excitatory neuron. The spike rate is expressed as a fraction of the maximal rate. In the following we will study two cases for the transduction function:

- Threshold linear, with threshold $\theta_e$ and gain $g_e$:

$$\phi(x) = \begin{cases} 1 & \text{if } x > \theta_e + 1/g_e \\ g_e(x - \theta_e) & \text{if } \theta_e < x < \theta_e + 1/g_e \\ 0 & \text{if } x < \theta_e \end{cases} \tag{3}$$

- Noisy integrate-and-fire, obtained by assuming a leaky integrate-and-fire neuron with threshold $\theta_e$, integration time constant of soma depolarization $\tau$ and absolute refractory period $\tau_0$, receiving a background stationary gaussian current of mean $\mu$ and RMS $\sigma$ [8].

Both transfer functions are shown in Fig. 1. The first one has the advantage of making possible the calculation of closed expressions for the delay activities as a function of the network parameters, while the second one is more realistic. We find that both functions lead to qualitatively similar results for the statistical properties of the attractors.

## 2.2 The inhibitory response

A separate inhibitory network is composed of $N_{inh}$ inhibitory neurons. The inhibitory reaction is unstructured, and every inhibitory neuron is connected to and from all excitatory neurons with a uniform synaptic strength. Every inhibitory neuron receives the same current from the excitatory neurons, and the entire inhibitory network becomes equivalent to a single inhibitory neuron characterized by its afferent current $I_{inh}(t)$ and its spike rate $V_{inh}$ (see e.g. [12]). The dynamics of the synaptic current into this inhibitory neuron is given by

$$\tau_{inh}\dot{I}_{inh} = -I_{inh} + J_{inh}\sum_j V_j \tag{4}$$

and

$$V_{inh} = \phi_{inh}(I_{inh}) \tag{5}$$

where $\phi_{inh}$ is the inhibitory current-to-rate transduction function.

In other words, the inhibitory neuron is driven by a current proportional to the mean activity in the excitatory network. The inhibitory response is the same to every excitatory neuron: for all $i$ $T_i$ in Eq. (1) is given by

$$T = V_{inh} = \phi_{inh}(I_{inh}). \tag{6}$$

Eqs. (1, 4, 6) together with the transduction functions $\phi_{exc}$ and $\phi_{inh}$ describe the dynamics in full, once the synaptic matrix connecting the excitatory neurons, $J_{ij}$, and $J_{inh}$ are given.

In the following we will restrict ourselves to a threshold-linear inhibitory transduction function. We have checked that using a function of the 'integrate-and-fire' type leads to the same qualitative results. This transfer function is characterized by the gain $g$ and the threshold $\theta$

$$\phi_{inh}(x) = \begin{cases} g(x-\theta) & \text{if } x > \theta \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

## 2.3 The learned synaptic matrix

The synaptic matrix reflects the learning of a sequence of $p$ binary patterns (stimuli) $(\{\eta_i^\mu = 0, 1\},\ i = 1, \ldots, N,\ \mu = 1, \ldots, p)$ presented in a fixed order during the training session [2, 5]. Each stimulus activates $fN$ neurons in the excitatory network, i.e. for all $\mu$

$$\sum_i \eta_i^\mu = fN,$$

where $f$ is the coding level of the stimuli. The population of active neurons in pattern $\mu$, i.e. $\{i|\eta_i^\mu = 1\}$, is denoted $F_\mu$. The sequence is considered periodic, i.e. pattern number $p+1$ is identical with pattern number 1. The binary form of the stored memories may be considered as a symbolic indication of the neurons whose rates are elevated by the stimulus and those which are not. The rates maintained by the neurons in the network are analog. We model the synaptic matrix following Willshaw [10], and the synaptic efficacies take only three values. The elements of the synaptic matrix $J_{ij}$ are taken to be:

$$J_{ij} = \frac{J}{fN}$$

if there exists at least one pattern $\mu$ for which $\eta_i^\mu = \eta_j^\mu = 1$ (except for the normalization this is the original Willshaw prescription);

$$J_{ij} = \frac{aJ}{fN}$$

where $0 < a < 1$, if $J_{ij} = 0$, by way of the Willshaw prescription, and there exists at least one pair of consecutive patterns $\mu, \mu + 1$ for which $\eta_i^\mu = \eta_j^{\mu+1} = 1$ or $\eta_j^\mu = \eta_i^{\mu+1} = 1$; otherwise

$$J_{ij} = 0.$$

The parameter $a$ represents the strength with which the contiguity of two patterns in the learned sequence is imprinted during learning (see e.g. [5]).

The self excitatory coupling is set to zero, $J_{ii} = 0$. The excitatory to inhibitory synaptic strength $J_{inh}$ is set to $J_{inh} = 1/fN$. Here we have restricted ourselves to a synaptic matrix which is prescribed a priori. A matrix which performs in a similar way can be generated also by a quasi-realistic learning process. This process will be discussed elsewhere [11].

# 3 Dynamics for non-overlapping memories

In this section we analyse the dynamics and determine the delay activities of neurons and their statistical properties in simplified conditions. The simplifications are:

- The patterns in the sequence of stimuli have no foreground neurons in common. This is a good approximation only in the limit of a very low coding level;
- The inhibition is taken to be instantaneous, i.e. the inhibitory currents afferent on any excitatory neuron is determined by the mean activity in the excitatory network at the same time;
- In section 3.3 all gain functions are taken to be threshold-linear.

Since in the experiment the typical values of $f$, the fraction of neurons activated by a stimulus, is low (about 1%), the probability that a neuron is activated by two given patterns is in fact low. Hence, to obtain an approximate idea of the structure of the delay activities we consider the case in which each neuron is in the foreground of at most one single pattern. Thus the populations $F_\mu$ of neurons activated by stimulus $\mu$ are non-overlapping sets of neurons. If the total number of memorized patterns is $p$, then this implies that $fp < 1$.

## 3.1 Dynamics of the overlaps

We define the mean activity among neurons in the foreground of pattern number $\mu$, $F_\mu$, by

$$m_\mu(t) = \frac{1}{fN} \sum_{j \in F_\mu} V_j(t) = \frac{1}{fN} \sum_j \eta_j^\mu V_j(t), \tag{8}$$

in which $\eta_j^\mu = 1$ defines the foreground group of pattern $\mu$, see e.g. section 2.3. In the case of non-overlapping patterns the mean total activity of the excitatory neurons $A(t)$ is

$$A(t) = \frac{1}{fN} \sum_i V_i = \sum_\mu m_\mu(t). \tag{9}$$

The excitatory synaptic current into neuron $i$ is given by

$$h_i(t) = \sum_j J_{ij} V_j(t) = \sum_{j,\mu} J_{ij} \eta_j^\mu V_j(t) \tag{10}$$

for non-overlapping patterns. The synaptic efficacies of section 2.3 can in this case be divided in three distinct classes:

- If both neurons $i$ and $j$ belong to the same population $F_\mu$ the synaptic efficacy is $J_{ij} = J/fN$;
- If neurons $i$ and $j$ belong to populations corresponding to two neighbouring stimuli, i.e. $F_\mu$ and $F_{\mu\pm1}$, $J_{ij} = aJ/fN$;
- Otherwise $J_{ij} = 0$.

If $i \in F_\mu$ the sum on the right hand side of Eq. (10) can be divided in three contributions corresponding to $F_{\mu-1}$, $F_\mu$ and $F_{\mu+1}$ i.e.

$$h_i(t) = \sum_j J_{ij} \eta_j^{\mu-1} V_j(t) + \sum_j J_{ij} \eta_j^\mu V_j(t) + \sum_j J_{ij} \eta_j^{\mu+1} V_j(t)$$

and all neurons belonging to a given population $F_\mu$ receive the same afferent excitatory current $h_\mu$ which is

$$h_\mu(t) = J m_\mu(t) + aJ[m_{\mu-1}(t) + m_{\mu+1}(t)]. \tag{11}$$

Consequently, the frequencies of neurons belonging to that population, become equal after a transient period, and by Eq. (8)

$$V_i = V_\mu = m_\mu \tag{12}$$

if $i \in F_\mu$.

The current dynamics of the excitatory neurons and the single inhibitory neuron can be expressed in terms of the mean activities $m_\mu(t)$, of neurons in the foreground set $F_\mu$ of each of the learned patterns. Following the removal of the stimulus ($H_i^{ext} = 0$ for all $i$) the network evolves according to:

$$\tau_{exc}\dot{I}_\mu = -I_\mu + Jm_\mu + aJ[m_{\mu-1} + m_{\mu+1}] - T, \tag{13}$$

$$m_\mu = \phi_{exc}(I_\mu), \tag{14}$$

$$\tau_{inh}\dot{I}_{inh} = -I_{inh} + \sum_\mu m_\mu, \tag{15}$$

$$T = \phi_{inh}(I_{inh}). \tag{16}$$

## 3.2 Dynamics with instant inhibition

As a further simplification the reaction of the inhibitory neurons is taken to be instantaneous: at each time $I_{inh}(t) = A(t)$ given by Eq. (9), i.e. the total activity of the excitatory network at the same time. As a result, the current received by an excitatory neuron, belonging to the group $F_\mu$, evolves according to

$$\tau_{exc}\dot{I}_\mu = -I_\mu + Jm_\mu + aJ[m_{\mu-1} + m_{\mu+1}] - \phi_{inh}(A) \tag{17}$$

with $A(t)$ given in terms of the $m_\mu$s by Eq. (9). Since there is now a single time constant we set $\tau_{exc}$ as the time unit, i.e. $\tau_{exc} = 1$. The effect of the stimulus is represented by an initial set of rates $V_i(t = 0)$ and hence by an initial set of $m_\mu(t = 0)$'s. Then, for a given pair of transfer functions and a specific value of $a$, one can solve these equations and study the correlations between attractors, when the network reaches one. The persistent delay activities are those activity distributions whose corresponding $m_\mu$'s make the r.h.s of Eq. (17) vanish. Their excitation depends only on the particular stimulus presented and not on the order in which stimuli had been learned (see e.g. [2]).

## 3.3 Explicit solution for threshold-linear transfer functions

A complete explicit solution for the attractors of the dynamics Eqs. (13, 14, 15) is possible if both gain functions are threshold-linear [9]. We will see in the next subsection that taking more realistic transduction functions produces qualitatively similar results. The transfer functions are given by Eqs. (3, 7), with $\theta = 1$, and we set $J = 1$.

The initial conditions are: $m_\nu(0) = m_0$, and $m_\mu(0) = 0$ for $\mu \neq \nu$. This corresponds to the situation in which stimulus $\nu$, one of the uncorrelated stimuli learned in the sequence, has been imposed on the (overtrained) network by the external currents.

### 3.3.1 Dynamics for $a = 0$

In this case the initial uncorrelated pure patterns are stable points of the dynamics, i.e. there is a fixed point solution

$$m_\nu = 1$$
$$m_{\nu \pm k} = 0 \quad \text{if } k \geq 1.$$

provided $g_e > 1$, and $0 < m_c < 1$ where

$$m_c = \frac{\theta_e g_e}{g_e - 1}.$$

For the parameters taken in Fig. 1, one has $m_c = 0.5$. It is shown in Appendix A that any configuration with $m_\nu(0) > m_c$ converges to the attractor, and that it is stable under small perturbations. This is also a stable attractor for small values of $a$, i.e. if $a < \theta_e$ the pure patterns are still stable points of the dynamics.

### 3.3.2 Dynamics for $a > 0$

In the following we set $\theta_e = 0$, $g_e = 1$. The dynamics of the network depends now on the parameters $a$ and $g$, the inhibitory gain. For any positive $a$, the population $F_\nu$ of neurons which are activated by the stimulus will activate neurons of populations $F_{\nu\pm1}$, which may in turn activate neurons from population $F_{\nu\pm2}$, etc. In this way excitation could percolate to all neurons. What prevents the propagation of excitation to a large number of populations of neurons is the inhibition, and the balance between the parameters $a$ and $g$ will control how many of these populations will be active in an attractor. In the next sections we present the results and their dependence on the ratio $a/g$. Only the delay activities are presented. The analysis of the dynamics of the network is presented in Appendix B.

We show in Fig. 2 the different regions (a phase diagram) in the $a$–$g$ plane. The parameter $K$ indicates the maximal distance $k$ at which a population $F_{\mu\pm k}$ gets activated when stimulus number $\mu$ is shown to the network. For example if $K = 1$ $F_{\mu-1}$, $F_\mu$ and $F_{\mu+1}$ are active in the attractor after presentation of stimulus $\mu$. In general, the number of populations activated in an attractor is $2K + 1$. Decreasing the inhibitory gain one can arrive at arbitrarily high values of the parameter $K$, whatever the value of the parameter $a$. As $K$ increases neurons from a larger number of pure stimuli participate in each attractor (delay activity) and hence attractors with larger separation become correlated. In this sense the range of correlations found in the experiment is essentially a measure of the inhibitory gain.

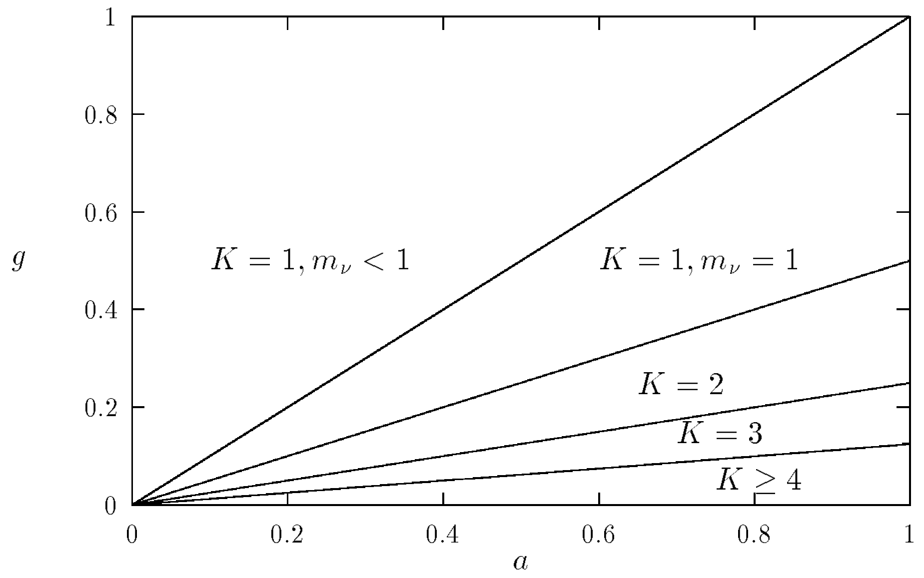


Figure 2: Different regions in the $a$–$g$ plane. $a$ represents the strength with which the contiguity of two patterns is imprinted and $g$ is the inhibitory gain. Each region is characterized by the parameter $K$ which characterizes the number of learned stimuli mixed into the delay activity distribution (see text).

#### 3.3.3 Attractors for high inhibitory gain $g > a$

The only non-zero fixed point solutions, symmetric around $\nu$, are

$$m_\nu = \frac{g}{2g - a}, \tag{18}$$

$$m_{\nu\pm1} = \frac{1}{2} m_\nu, \tag{19}$$

$$m_{\nu\pm k} = 0 \quad \text{if } k \geq 2. \tag{20}$$

Thus for small positive $a$ and $g \gg a$ we get $m_\nu \sim 0.5$, $m_{\nu\pm1} \sim 0.25$. The solution is discontinuous at $a = 0$.

#### 3.3.4 Attractors for low inhibitory gain $g < a$

We denote $K = E(a/2g) + 1$ where $E(x)$ is the integer part of $x$, and $\lambda = a/2g - E(a/2g)$ with $0 < \lambda < 1$. The only fixed point solution symmetric around $\nu$ is case:

$$m_{\nu\pm k} = 1 \quad \text{if } k \leq K - 1 \tag{21}$$

$$m_{\nu\pm K} = \lambda \tag{22}$$

$$m_{\nu\pm k} = 0 \quad \text{if } k \geq K + 1 \tag{23}$$

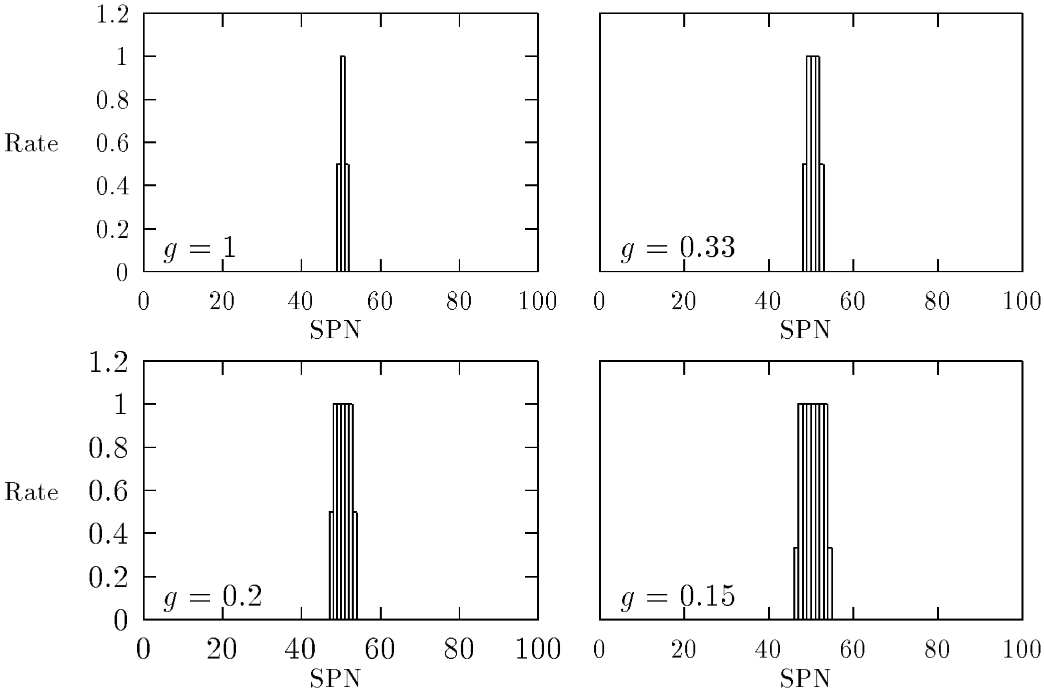


Figure 3: Delay activity *vs* serial position number for a specific neuron which is in the foreground of pattern number 50. A bar represents the delay activity after presentation of the corresponding stimulus. Rate is expressed as a fraction of saturation rate. Parameters of the network: $p = 100$, $a = 1$, and $g$ is indicated on the figures.

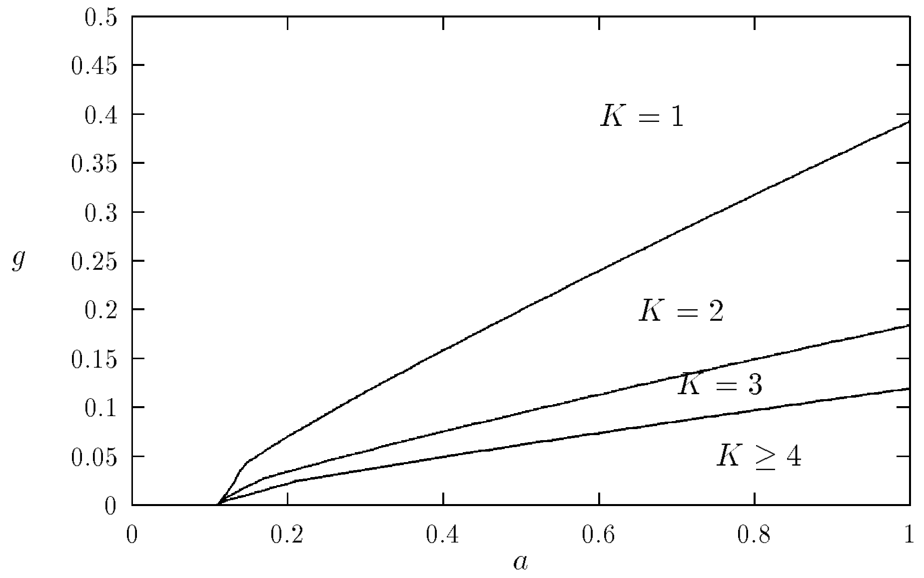


Figure 4: Phase diagram in the $a$-$g$ plane, for the 'integrate-and-fire' transfer function. Compare with Fig. 2.

Thus $K$ is the most distant pattern in the sequence with which the attractor has a non-zero overlap. In other words, it is the most distant stimulus whose foreground neurons are active in the attractor.

We show in Fig. 3 the delay activities of a neuron in different attractors as a function of the serial position number (SPN) of the shown stimulus. The chosen neuron is in the foreground of stimulus number 50. As the inhibitory gain $g$ is decreased, this neuron is activated in an increasing number of neighboring attractors. This figure is qualitatively similar to Fig. 3b of Ref. [2]. Note however that in Fig. 3 the mean activity of the neuron is at saturation for all but the most distant attractor in which this neuron has nonvanishing delay activity, contrary to the experimental figures. This is due to the particular shape of the transduction function.

## 3.4 'Integrate-and-fire' transfer functions

In this case we iterate numerically the recurrence equations (13-15) until the fixed point is reached. The results are qualitatively similar as in the previous section. Decreasing the gain of the inhibitory response function, the attractors become correlated with more patterns in the sequence. As an example we take the excitatory transfer function shown in Fig. 1, keeping a threshold-linear inhibitory transfer function, as in Eq. (7) with threshold $\theta = 0.05$.

The 'phase diagram' in this case is shown in Fig. 4. It is very similar to the phase diagram of Fig. 2, obtained for a threshold-linear transduction function.

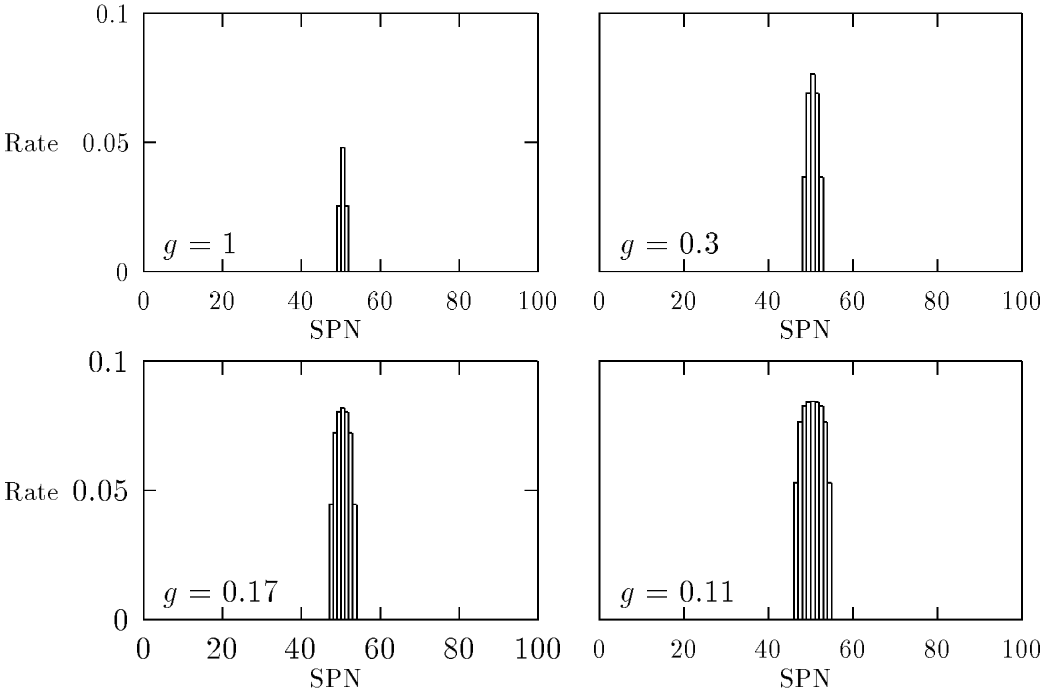


Figure 5: Delay activity *vs* serial position number histogram for a specific neuron which is in the foreground of pattern number 50. Parameters of the network: $p = 100$ stored stimuli, $a = 1$, $\theta = 0.05$. $g$ is indicated on the figures. This figure is the analog of Fig. 3 for an excitatory integrate-and-fire transduction function. Note that in this case the neuron is always much below saturation.

The main difference is that there is a critical value of $a$, $a \sim 0.11$, under which $K = 1$ for any value of the inhibitory gain $g$. This means that in this case $a$ is too low to provoke activation of populations which are not the nearest neighbors of the population corresponding to the presented stimulus. For $a > 0.11$, one has a situation similar to the case of threshold-linear transfer function, and one can obtain an arbitrarily high value of $K$ decreasing the inhibitory gain $g$. Another difference with the threshold-linear case is that there is no discontinuity at $a = 0$ — the delay activity of the next neighbour population increases continuously from zero as $a$ increases.

The rates versus the serial position number (as in Fig. 3), for a specific neuron which is in the foreground of pattern number 50, are shown in Fig. 5 for $p = 100$ stored memories, $a = 1$ and different values of the inhibitory gain. Stimuli are shown with $m_\nu = 0.1$. These figures are again the analogs of Figs. 3a and 3b of Ref. [2]. The neuron is only active when neighbours of stimulus number 50 are presented to the network. This 'clustering' of the neuronal response along the SPN is similar to the experimental figures. When the inhibitory gain $g$ goes down the cluster includes more patterns which are neighbors of pattern number 50. Note that the rates cluster around the SPN of the stimulus for which the neuron is in the foreground as in the case of threshold-linear gain functions. The positive rates are now more realistic. When the inhibitory gain $g$ decreases, such a neuron, as well as all selective neurons, is active in the delay activity period following the presentation of more separated neighbors in the sequence.

To test the robustness to the form of the inhibitory response function we tried the same type of gain function as for the excitatory neuron. We took for the inhibitory gain function, $\phi_{inh}$ in Eq. (5), the same function as the excitatory gain function $\phi_{exc}$ in Fig. 1, i.e. $\phi_{inh} = g\phi_{exc}$. The results remain qualitatively unchanged.

# 4 Statistical properties of the attractors

## 4.1 Standard correlations

### 4.1.1 Definitions

The delay activity of neuron $i$ in the attractor provoked by stimulus $\mu$ is denoted by $V_i^\mu$. We consider the distribution of mean rates $V_i^\mu$ in a sample $\mathcal{S}$ of $N_s$ neurons as a random variable. Its mean is

$$\overline{V_\mu} = \frac{1}{N_s} \sum_{i \in \mathcal{S}} V_i^\mu, \tag{24}$$

its variance is

$$\overline{\Delta V_\mu^2} = \frac{1}{N_s} \sum_{i \in \mathcal{S}} (V_i^\mu)^2 - \overline{V_\mu}^2 \tag{25}$$

and the covariance of a pair of such random variables, corresponding to a given pair of attractors $\mu$ and $\nu$ is

$$\mathrm{Cov}_{\mu\nu} = \frac{1}{N_s} \sum_{i \in \mathcal{S}} V_i^\mu V_i^\nu - \overline{V_\mu}\,\overline{V_\nu}$$

The correlation between the activity distributions in the two attractors $\mu$ and $\nu$ is

$$C_{\mu\nu} = \frac{\mathrm{Cov}_{\mu\nu}}{\sqrt{\overline{\Delta V_\mu^2}\;\overline{\Delta V_\nu^2}}}. \tag{26}$$

The mean correlation between two attractors at distance $k$ is defined as

$$C_k = \frac{1}{p} \sum_{\mu} C_{\mu,\mu+k}, \tag{27}$$

where $p$ is the total number of memorized attractors.

#### 4.1.2 Correlations between attractors

For non-overlapping patterns the usual correlations, defined in the last section, are easily calculated from the knowledge of the delay activities in terms of the $m_\mu$. When stimulus $\mu$ is presented the delay activity of a neuron $i \in F_\nu$ is $V_i^\mu = m_\nu^\mu$. Thus $m_\mu^\nu$ is the delay activity of population $\nu$ after stimulus $\mu$ is presented, and is given by Eqs. (21-23) or (18-20) depending on the value of the parameters $a$ and $g$. The mean rate averaged over a sample $\mathcal{S}$ is

$$\overline{V_\mu} = \frac{1}{N_s} \sum_{i \in \mathcal{S}} V_i^\mu = \frac{1}{N_s} \sum_{\nu} m_\nu^\mu N_s^\mu,$$

where $N_s^\mu$ is the number of neurons in the sample $\mathcal{S}$ which are in the foreground of pattern $\mu$. If the sample of recorded neurons is large enough and contains neurons belonging to the different populations $F_\mu$ in equal parts, we have for all $\nu$

$$\frac{N_s^\mu}{N_s} \sim f$$

and thus

$$\overline{V_\mu} = f \sum_{\nu} m_\nu^\mu.$$

Similarly we obtain for the variance $\overline{\Delta V_\mu^2}$ and covariance $\mathrm{Cov}_{\mu,\mu+k}$

$$\overline{\Delta V_\mu^2} = f \sum_{\nu} (m_\nu^\mu)^2 - f^2 \left( \sum_{\nu} m_\nu^\mu \right)^2$$

$$\mathrm{Cov}_{\mu,\mu+k} = f \sum_{\nu} m_\nu^\mu m_{\nu+k}^\mu - f^2 \left( \sum_{\nu} m_\nu^\mu \right)^2$$

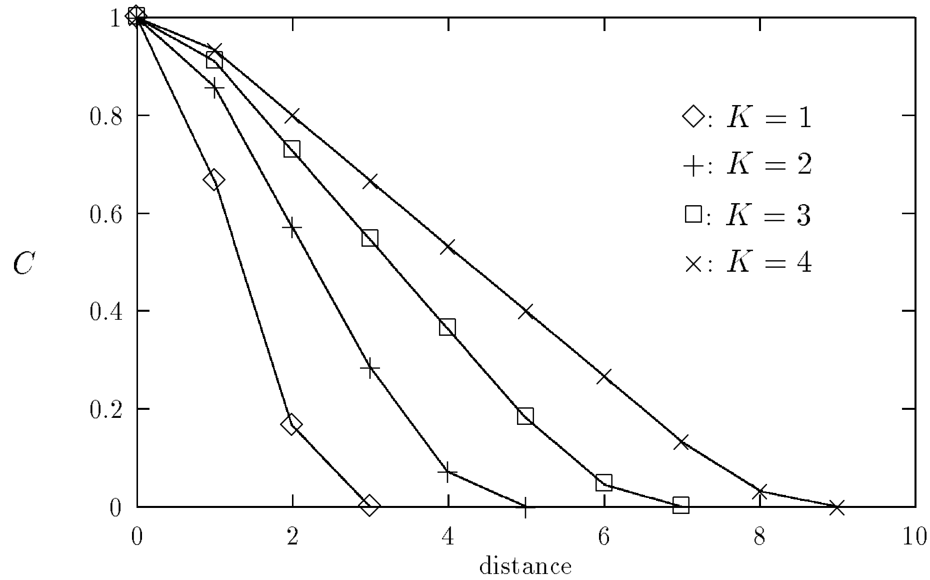


Figure 6: Standard correlations between attractors vs distance between the corresponding stimuli (Eqs. 28-30, 31-33) in the sequence, for several values of $K$ (indicated in the figure), and $\lambda = 0.5$, for threshold-linear transfer functions.

All these expressions are independent of the index $\mu$, since in the present simplified model the vectors of $m_\mu^\nu$s describing attractors corresponding to different stimuli have the same components and are simply translations of each other. The index $\mu$ is therefore dropped. Finally, if the coding level $f$ is low, the terms in $f^2$ in Eq. (26, 27) can be neglected and we obtain for the correlation of delay activities at separation $k$ in the training sequence

$$C_k = \frac{\sum_\nu m_\nu m_{\nu+k}}{\sum_\nu m_\nu^2} + O(f)$$

For a high inhibitory gain, $a < g$, the correlations between attractors at distance $k$ are:

$$C_1 = \frac{2}{3} + O(f) \tag{28}$$

$$C_2 = \frac{1}{6} + O(f) \tag{29}$$

$$C_k = 0 \quad \text{if } k \geq 3. \tag{30}$$

This is the leftmost curve in Fig. 6.

For lower inhibitory gain, $g < a$, the correlations between attractors are

$$C_k = \frac{2K - 1 - k + 2\lambda}{2K - 1 + 2\lambda^2} + O(f) \quad 1 \leq k \leq 2K - 1 \tag{31}$$

$$C_{2K} = \frac{\lambda^2}{2K - 1 + 2\lambda^2} + O(f) \tag{32}$$

$$C_k = 0 \quad k \geq 2K + 1 \tag{33}$$

In Fig. 6 we plot these correlations for $K = 2, 3, 4$ and $\lambda = 0.5$. The correlations decrease from 1 to zero when the distance between the attractors increases. The distance at which the correlations become zero, $2K + 1$, is the number of different populations involved in a given attractor. As $K$ increases the correlations between attractors corresponding to neighbouring stimuli become very high — for example $C_1 \sim 0.93$ for $K = 4$. In this case the correlations also have a long tail, up to a distance $k = 8$.

## 4.2 Kendall rank coefficients (KRCs)

KRCs, rather than usual correlation coefficients (Sec. 4.1), of delay activity distributions have been used to represent the results in the experiments of Miyashita [2]. We proceed to discuss, in detail, the corresponding exact KRCs for the case of non-overlapping stimuli and then the approximate KRCs for uncorrelated patterns with low coding level. The exact KRCs are contrasted with the corresponding conventional correlations.

### 4.2.1 Definitions

KRCs are calculated independently for each neuron. For a given neuron one computes

$$U^k_{\mu\nu} = \mathrm{sign}[(V_i^\mu - V_i^\nu)(V_i^{\mu+k} - V_i^{\nu+k})] \tag{34}$$

with $k$ $(1 \leq k \leq p/2)$ fixed, where $V_i^\mu$ is the mean activity of neuron $i$ in attractor $\mu$. The KRC of neuron $i$ at distance $k$ is the mean value of the elements of the matrix $U^k$, i.e.

$$R_k = \frac{2}{p(p-1)} \sum_{\mu<\nu} U^k_{\mu\nu} \tag{35}$$

See e.g. Ref. [14]. The idea of these correlation coefficients is to take into account not the values of the delay activity of a neuron, but the rankings between such values. The KRCs are then averaged over the sample of recorded neurons.

## 4.3 KRCs — non overlapping patterns

Consider, as an example, the case of non overlapping patterns and neurons with threshold-linear transfer functions, in the region $g < a < 2g$ ($K = 1$ in Fig. 2). When stimulus number $\nu$ is presented, the network relaxes to an attractor described by Eqs. (21-23) where $K = 1$. In that attractor neurons have elevated spike rates only if they belong to the foreground of one of three stimuli — numbered $\nu - 1$, $\nu$,

$\nu+1$. Conversely, if neuron $i$ belongs to the foreground of stimulus number $\rho$ it will be active only when stimuli $\rho-1$, $\rho$ or $\rho+1$ are presented, and the corresponding delay activities will be $V_i^\rho = 1$, $V_i^{\rho+1} = V_i^{\rho-1} = \lambda$ where $0.5 < \lambda < 1$.

In this situation, as an example we show $U^1_{\mu\nu}$ defined in Eq. (34)

$$
\begin{array}{c|cccccccccc}
 & & & & & & (\rho) & & & & \\
 & 0 & \cdots & & 0 & 1 & 1 & 0 & 0 & \cdots & 0 \\
 & \cdots & & & \cdots & \cdots & \cdots & \cdots & & & \cdots \\
 & & & & & 1 & 1 & 0 & & \cdots & 0 \\
 & 0 & \cdots & & 0 & 1 & 0 & -1 & 0 & \cdots & 0 \\
 & 1 & \cdots & 1 & 1 & 0 & -1 & 0 & 1 & \cdots & 1 \\
(\rho) & 1 & \cdots & 1 & 0 & -1 & 0 & 1 & 1 & \cdots & 1 \\
 & 0 & \cdots & 0 & -1 & 0 & 1 & 0 & & \cdots & 0 \\
 & & & & 0 & 1 & 1 & & & \cdots & 0 \\
 & \cdots & & & \cdots & \cdots & \cdots & \cdots & & & \cdots \\
 & 0 & \cdots & & 0 & 1 & 1 & 0 & & \cdots & 0
\end{array}
\qquad (36)
$$

Here we take the elements of the matrix to be zero if

$$(V_i^\mu - V_i^\nu)(V_i^{\mu+k} - V_i^{\nu+k}) = 0.$$

The measured delay activity of a real neurons is, of course, never strictly zero due to the spontaneous activity. However, if the activity of the neuron after presentation of stimuli $\mu, \nu, \mu+k, \nu+k$ is due to spontaneous activity we may assume that it is random and uncorrelated and hence that $U^k_{\mu\nu}$ will be 1 or $-1$ with equal probability. The average over all elements of the matrix would be the same as if these random elements were taken to be zero. The standard deviation of the resulting KRC, would be proportional to the square root of the number of such elements in the matrix, i.e of order $\sqrt{2}/p$ for large $p$ and small $f$.

The sum of the elements of the matrix of Eq. (36) is $4(p-4)$, and their mean, Eq. (35), is

$$R_1 = \frac{4(p-4)}{p(p-1)}, \qquad (37)$$

which is the mean KRC of one neuron.

In the model network selective neurons, i.e. neurons which have an elevated activity in response to at least one stimulus, are neurons which belong to the foreground of one pattern in the sequence. All selective neurons in this case have equal KRCs. We will see in the next section that this does not hold when random, uncorrelated patterns are considered. Moreover, in the present case non selective neurons have all KRCs equal to zero. However, to allow a comparison with Fig. 3c of Ref. [2] these neurons should be discarded in calculating the mean value of the KRCs, since all data bases considered are composed of selective neurons. Hence the mean of $R_1$ over any sample of selective neurons is just $R_1$ of Eq. 37. Here we encounter an instance in which the KRCs vary with the number of learned patterns while the relative structures of the delay activities remain unchanged, Eq. (21-23). See e.g. [7].

The complete description of the KRCs in the different parameter regions of Section 3.2 is:

- **Region $K = 1$ in Figs. 2, 4**: only three populations are active in the attractor following each presentation. The rank correlation coefficients for selective neurons are:

$$R_1 = \frac{4(p-4)}{p(p-1)} \tag{38}$$

$$R_2 = \frac{2(p-11)}{p(p-1)} \tag{39}$$

$$R_k = -\frac{18}{p(p-1)} \quad \text{if } k \geq 3. \tag{40}$$

Thus for large $p$ $R_1 \sim 4/p$, $R_2 \sim 2/p$, $R_k \sim -18/p^2$. This is to be compared with the correlations of Eqs. (28-30) or (31-33).

- **Region $K > 1$ in Figs. 2, 4**: the attractors involve populations up to a distance $K$, where $K = E(a/2g) + 1$. The matrix $U_k$ has now a group of $2K + 1 - k$ columns and rows with non-zero elements. For large $p$ the mean of the elements of the matrix will be dominated by the elements that do not belong to the intersections between these columns and rows. These elements are equal to 1. Thus the rank correlation coefficient are, for large $p$ and $k \leq 2K$

$$R_k \sim \frac{2(2K+1-k)}{p} \tag{41}$$

When $k > 2K$ the columns and rows vanish and only squares of $(2K+1)^2$ $-1$ elements remain. The mean becomes negative and

$$R_k \sim -\frac{2(2K+1)^2}{p^2} \quad \text{for } k > 2K \tag{42}$$

These values should be compared with the correlations of Eq. (31-33).

For large $p$ the KRCs become a linear decreasing function of $k$, going from $R_1 = 4K/p$ to $R_{2K+1} = -O(1/p^2)$. For high $k$ the coefficients are always negative.

The KRCs are shown in Fig. 7 for different values of $K$, and $p = 100$. At this value of $p$ the KRCs decrease almost linearly with the distance between the corresponding stimuli, until they reach a negative value at distance $2K + 1$. Note that the values of the rank coefficients depend not only on the distance $K$, but also on the number of memorized patterns $p$. They are independent of the ratio between the different non-zero delay activities in a given attractor. The usual

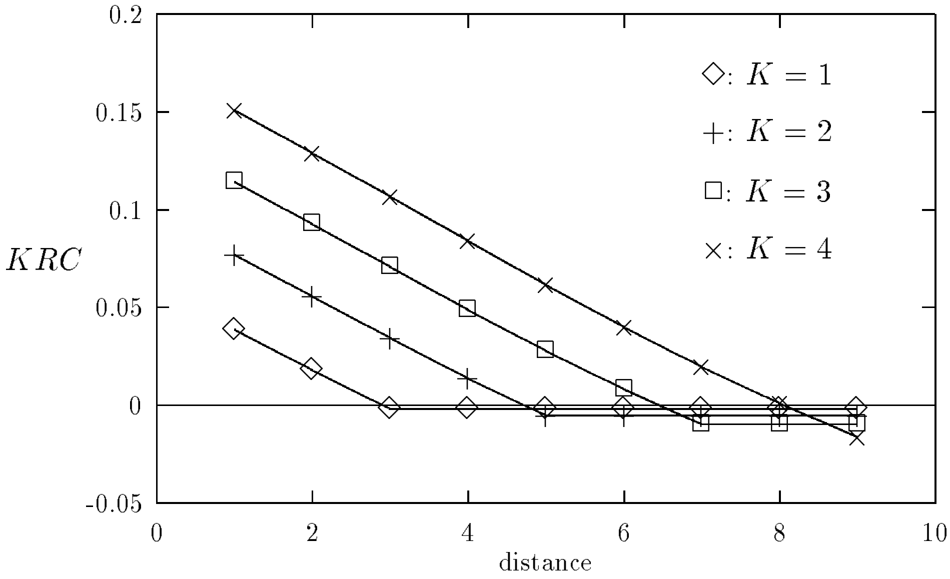


Figure 7: KRCs of delay activity distributions as a function of the distance between the corresponding stimuli in the sequence, for $p = 100$ and different values of $K$, indicated in the figure.

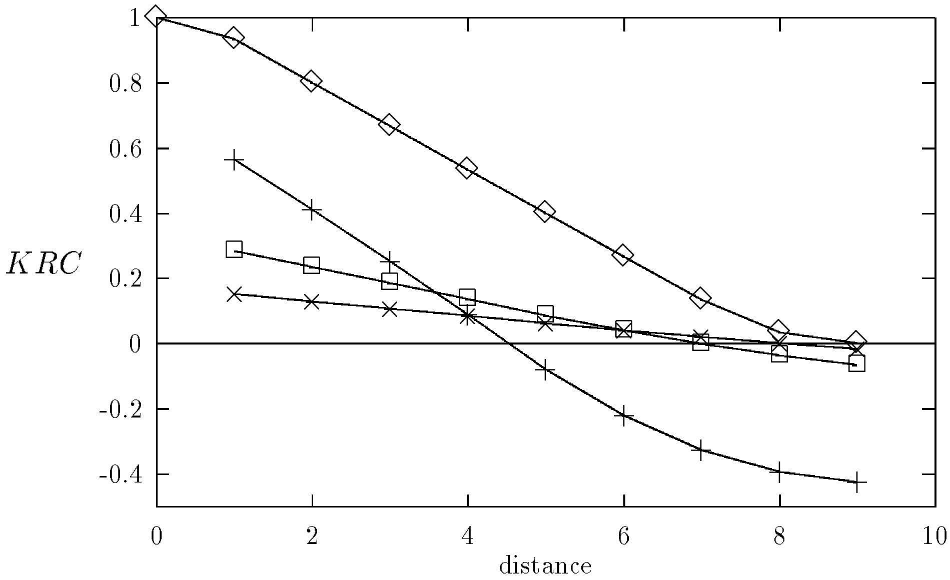


Figure 8: Comparison of correlation coefficients and Kendall rank coefficients for the same network dynamics. Diamonds: usual correlations. $+$: KRCs for $p = 20$. $\square$: KRCs for $p = 50$. $\times$: KRCs for $p = 100$. The usual correlations go continuously from one to zero with increasing separation, and are independent of the number of patterns. The KRCs at every separation depend on $p$, and decrease to a negative asymptotic value.

correlations are independent of $p$. This difference is depicted in Fig. 8. We plot both KRCs and usual correlation coefficients for a network with threshold-linear transfer functions, with parameters given in Fig. 5 and $g = 0.15$, for three values of $p$. For these parameters we have $K = 4$. The usual correlation coefficients decrease from $C_0 = 1$ to $C_{2K+1} = C_9 = 0$ independently of the number of patterns, while the KRCs are quite sensitive to the value of $p$. KRCs become negative after some value depending on $p$. They become constant at the same separation at which the usual correlations vanish, i.e. $2K + 1$, for all $p$. The dependence on $p$ of the KRCs is peculiar to the special way in which they are defined, and is not descriptive of the dynamics. For the same network, choosing sets with a different number of attractors will lead to similar conventional correlations, but to very different KRCs. Conventional correlations should therefore be preferred to represent the statistical properties of the delay activities.

## 4.4 Approximate KRCs — uncorrelated patterns

Until this section we have considered the situation in which every neuron in the network is activated by at most one stimulus. If patterns are randomly drawn with coding level $f$, there will be neurons activated by more than one stimulus. This is the situation we consider in this section. For the calculation of the coefficients we suppose that in every attractor neurons from an equal number $(2K + 1)$ of adjacent pure stimuli are active. We denote by $r$ the number of patterns for which the neuron under consideration is active. For this neuron, the matrix $U^k$, instead of having a single group of $2K + 1 - k$ non zero columns and rows, will have $r$ such groups. If the coding rate is very low then to a good approximation these groups of columns and rows in the matrix do not overlap. Hence, for large $p$, the rank coefficients will be, for $i \leq 2K$,

$$R_i \sim \frac{2r(2K + 1 - i)}{p} \tag{43}$$

For random uncorrelated patterns with coding rate $f$ the probability for a neuron being in the foreground of $r$ patterns is, for $0 \leq r \leq p$

$$p_r = \tbinom{p}{r} f^r (1 - f)^{p-r}$$

where $\binom{p}{r}$ are the binomial coefficients. The selective neurons are the ones that are activated by at least one pattern. The average of $r$ over these neurons gives

$$\langle r \rangle_{r>0} = \frac{\sum_{r>0} r p_r}{\sum_{r>0} p_r}$$

which yields

$$\langle r \rangle_{r>0} = \frac{fp}{1 - (1 - f)^p}$$

Thus the mean coefficients are, for $i \leq 2K$,

$$\langle R_i \rangle_{r>0} \sim \frac{2f(2K+1-i)}{1-(1-f)^p}$$

If the average is done on a restricted class of selective neurons, those which have a particularly high coefficient $R_1$, i.e. whose $r$ is greater than some number $r_0$, (see upper curve in Fig. 3c of ref. [2]) then we have

$$\langle R_i \rangle_{r>r_0} \sim 2f(2K+1-i)\gamma$$

where

$$\gamma = \frac{\sum_{r=r_0}^{p} r\binom{p}{r} f^r (1-f)^{p-r}}{fp \sum_{r=r_0}^{p} \binom{p}{r} f^r (1-f)^{p-r}}$$

Thus averaging over neurons which have a high first neighbour coefficient $R_1$ increases all positive coefficients with the same proportion. For example, if $r_0 = 1$ we get

$$\langle R_i \rangle_{r>1} \sim \frac{2f(1-(1-f)^{p-1})(2K+1-i)}{1-(1-f)^{p-1}(1+(p-1)f)} \tag{44}$$

For $i > 2K$, the rank coefficients will be for large $p$

$$R_i \sim -\frac{2r^2(2K+1)^2}{p^2} \tag{45}$$

From Eqs. (45) it follows that selecting neurons which have a high coefficient $R_1$ will also increase the absolute value of the negative coefficients.

The results of this analysis can be directly confronted with the experimental data presented in Fig. 3c of Ref. [2], as was done in Ref. [7]. If one supposes that the population of neurons used to calculate the 'high' KRCs in the experiment are neurons which are in the foreground of more than one stimulus, one can get an estimate of the coding level from the ratio of the number of these neurons to the number of selective neurons. In fact, these neurons with 'high' KRCs are about half of selective neurons (28 out of 57). If the stimuli are uncorrelated, this ratio is obtained when the coding level is $f \sim 0.0125$. Taking $K = 3$ as an estimate of the maximal distance at which populations of neurons are activated, both average and 'high' KRCs are of the same order of magnitude as the experimental ones. The comparison for both KRCs is shown in Fig. 9.

This analysis also suggests that it would be interesting to measure the distribution of the KRCs among neurons. Eq. (43) shows that the KRC of a given neuron is determined by the number of stimuli by which it is activated. This holds when the activity of a neuron which does not participate in an attractor is zero. As noticed in Sec. 4.3, in a real network we may assume that when a neuron does not participate in an attractor it has a random spontaneous activity. Because of these random uncorrelated delay activities, one has to add to Eq. (43) a noise term with

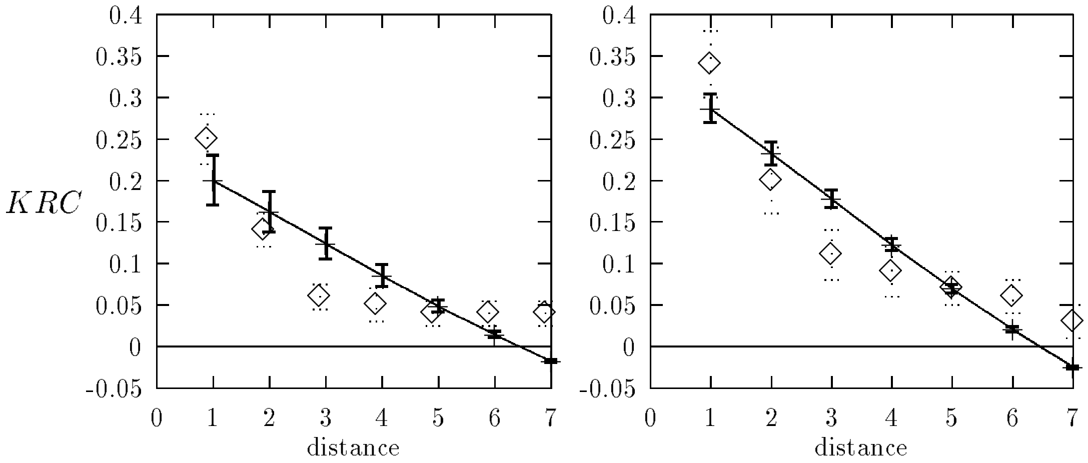


Figure 9: Comparison of analytical estimates of the KRCs for $K = 3$, $p = 100$ and $f = 0.0125$ with the experimental results of Miyashita (figure 3b of Ref. [2]). Left: average KRCs for selective neurons. $\diamond$: experimental data. $+$: analytical estimate. Right: average KRCs for neurons with high first neighbour KRC (see text). $\diamond$: experimental data. $+$: analytical estimate. The estimated error bars are standard errors, calculated for the same number of neurons as in the experiment, i.e. 57 selective neurons and 28 selective neurons with high first neighbour KRC.

zero mean and width $\sqrt{2}/p$. Thus the distribution of the KRCs among neurons which are activated by exactly $r$ stimuli will have a mean given by Eq. (43), and width $\sqrt{2}/p$. The resulting distribution of correlation coefficients will have peaks of decreasing weights, where the $n$th peak is centered at $\lambda n$, with $\lambda$ given, for coefficient at distance $i$ $R_i$, by

$$\lambda = \frac{2(2K+1-i)}{p},$$

and has a width of order $\sqrt{2}/p$ due to the randomnesss of the activity of neurons when they do not participate in an attractor. Peak number $n$ contains neurons which are in the foreground of $n$ different stimuli. The peaks at distance 1 are well separated, since the distance between peaks is large compared with the width of a single peak, which is of order $\sqrt{2}/p$. As the distance increases the peaks become closer till they merge together when their distance is of the order of the width of the peak (see e.g. Fig. 10). It would be interesting to have the experimental distribution of KRCs, since it would give an insight on the distribution of selective neurons.

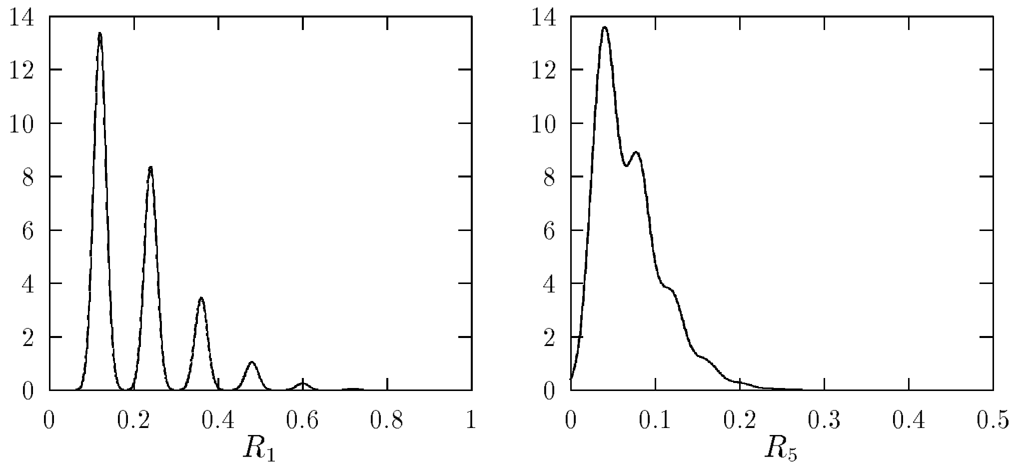


Figure 10: Distribution of KRCs obtained with $f = 0.0125$, $K = 3$ and $p = 100$. Left: distribution of $R_1$ — well separated peaks of width $\sqrt{2}/p$. Right: distribution of $R_5$ — peaks merge.

# 5 Conclusions and outlook

In this paper we have studied analytically the dynamics of an attractor neural network converting temporal correlations between learned stimuli into spatial correlations between attractors. The features of the neural delay activity versus the serial position number of the stimulus in the learned sequence, as well as statistical properties of the attractors, turn out to be quite similar to results of delayed match to sample experiments in inferotemporal cortex of monkeys. This detailed correspondence between the model and the experimental data carry much promise since the model is very simple and can be studied in detail.

For what concerns the details of the model, there are many unrealistic features like the prescribed synaptic matrix of section 2.3. However such a synaptic matrix can be replaced by a matrix obtained during a quasi-realistic learning dynamics [13, 11]. The availability of this learning procedure makes possible the study of the statistical properties of the attractors as a function of the presentation procedures. One can for example study the effect of sequences of ordered pairs of stimuli unordered among themselves. This situation corresponds to the experimental setup of Ref. [3]. The study of the learning dynamics leading to spatial correlations between attractors goes beyond the scope of this paper, and will be considered in a future publication [11]. Both excitatory and inhibitory neurons may also be oversimplified. Yet, in a complex environment as a live cortex it will be the assembly properties of the neural module which will determine how simplified the neural element is allowed to be. The behaviour of the network is very robust to changes in the excitatory and/or inhibitory transduction function. Thus one can expect the statistical properties of the delay activities not to depend much on the details of the model neuron, even if the resulting properties of the individual neuron, like the firing rate in the delay activity, are different.


## Acknowledgements

I wish to thank Daniel Amit for many discussions and a careful reading of the manuscript, and Marc Potters for many useful comments. A fellowship from programme Cognisciences of CNRS is gratefully acknowledged.


# Appendices

## A. Convergence to the fixed points, no correlations

In this section the case $a = 0$ is considered. If at $t = 0$ only one of the $m_\mu$s is nonzero, the dynamics (Eqs. 13-15) is completely characterized by the equation

$$\dot{m}_\nu = -m_\nu + \phi_{ex}(m_\nu)$$

Note that in this case if only one population is activated by the stimulus, inhibition does not play any role since it is kept below threshold. There are four different cases for the initial configuration $m_\nu(0) = m_0$.

- $m_0 < \theta_{ex}$: the activity dies down

$$m_\nu(t) = m_0 \exp(-t)$$

- $\theta_{ex} < m_0 < m_c$: one has, for

$$0 < t < t_0 = \frac{1}{g_{ex} - 1} \ln \frac{m_c - \theta_{ex}}{m_c - m_0},$$

$$m_\nu(t) = m_0 \exp[t(g_{ex} - 1)] + m_c(1 - \exp[t(g_{ex} - 1)]) \tag{46}$$

  where

$$m_c = \frac{\theta_{ex} g_{ex}}{g_{ex} - 1}.$$

  At $t = t_0$ one has $m_\nu = \theta_{ex}$. Then for $t > t_0$ one has

$$m_\nu(t) = \theta_{ex} \exp(t_0 - t)$$

  and the selective activity dies down, $m_\nu \to 0$.

- $\theta_{ex} + 1/g_{ex} > m_0 > m_c$: for

$$0 < t < t_1 = \frac{1}{g_{ex} - 1} \ln \frac{\theta_{ex} + 1/g_{ex} - m_c}{m_0 - m_c},$$

  $m_\nu(t)$ is given by Eq. (46) At $t = t_1$ one has $m_\nu = \theta_{ex} + 1/g_{ex}$. Then for $t_1 < t$ one has

$$m_\nu(t) = (\theta_{ex} + 1/g_{ex}) \exp(t_1 - t) + 1 - exp(t_1 - t)$$

  and the network converges to the attractor $m_\nu \to 1$.

- $\theta_{ex} + 1/g_{ex} < m_0$: in this case the network dynamics converge to the attractor $m_\nu = 1$

$$m_\nu(t) = m_0 \exp(-t) + 1 - exp(-t)$$

The stability of the selective fixed points is in this case guaranteed by the positive excitatory threshold: any small overlap with another memory will vanish exponentially.

## B. Convergence to the fixed points, correlations

In this section we take $\theta_{ex} = 0$, $g_{ex} = 1$, $\theta_{in} = 0$, $g_{in} = g$ and $0 < a \leq 1$. At $t = 0$ only one population, $F_\nu$, is activated, and $m_\nu(0) = m_0$ with $0 < m_0 \leq 1$. The dynamics is characterized by a sequence of times at which successive populations of neurons get activated or arrive at saturation. Starting from $t = 0$:

- $0 < t < t_0$: at $t = 0$ the populations $F_{\nu\pm1}$ get activated and thus $0 < m_{\nu\pm1} < m_\nu < 1$, while all other neurons (including the inhibition) stay silent. We have (from now we set $m_\nu = m_0$, and $m_{\nu\pm K} = m_K$)

$$\dot{m_0} = 2am_1$$

$$\dot{m_1} = am_0$$

  Thus both overlaps increase till $m_0 + 2m_1 = 1$ which occurs at $t = t_0$, at which time the inhibition is activated by the increase in the global activity of the network.

- $t_0 < t < t_1$: we still have $0 < m_1 < m_0 < 1$, but now inhibition is active and thus

$$\dot{m_0} = -gm_0 + 2(a - g)m_1 + g$$

$$\dot{m_1} = (a - g)m_0 - 2gm_1 + g$$

  If $a < g$, the network goes to an attractor in which $0 < m_1 < m_0 < 1$, where the overlaps are given by Eq. (18-20). Otherwise both overlaps increase till $m_0$ reaches saturation, $m_0 = 1$, at $t = t_1$.

- $t_1 < t < t_2$: $m_0 = 1 > m_1 > 0$. $m_1$ varies according to

$$\dot{m}_1 = -2gm_1 + a$$

If $a/2 < g < a$ the network goes to an attractor with $m_0 = 1$, and $m_1 = a/2g$. Otherwise $m_1$ increases till $m_2$ becomes positive, at time $t_2$.

- (...)

- $t_{2K+1} < t < t_{2K+2}$: $2K + 3$ populations are now active, $m_0 = \ldots = m_K = 1 > m_{K+1} > 0$. $m_{K+1}$ is given by

$$\dot{m}_{K+1} = -2gm_{K+1} + a - 2gK$$

If $a/(2K+2) < g < a/2K$ the network goes to an attractor with $m_0 = \ldots = m_K = 1$ and $m_{K+1} = a/2g + K$. Otherwise $m_{K+1}$ increases till $m_{K+2}$ becomes positive, at time $t_{2K+2}$.

- (...)

Thus in the case of a low inhibitory gain, $g < a$, the attractors are given by Eq. (21-23).

# References

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

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

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

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

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

[6] Amit DJ 1993 The Hebbian paradigm reintegrated: Local reverberations as internal representations, *BBS*, to be published.

[7] Amit DJ Brunel N and Tsodyks MV 1994 Correlations of cortical hebbian reverberations: theory vs experiment, *J. Neurosci.*, in press.

[8] 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

[9] Treves A 1990 Graded-response neurons and information encodings in autoassociative memories, *Phys. Rev.* **A42** 2418

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

[11] Brunel N 1994, Stochastic learning in networks converting temporal into spatial correlations, in preparation.

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

[13] Amit DJ and Fusi S 1994 Dynamic learning in neural networks with material sysnapses, *Neural Computation*, in press.

[14] Snedecor GW and Cochran WG 1969 *Statistical Methods* (Iowa State University, Ames, Iowa)