Adequate input for learning in attractor neural networks

Daniel J. Amit\textsuperscript{1} and Nicolas Brunel

INFN, Sezione di Roma, Istituto di Fisica
Università di Roma, La Sapienza, P.le Aldo Moro, Roma

Abstract

In the context of learning in attractor neural networks (ANN) we discuss the issue of the constraints imposed by the requirement that the afferents arriving at the neurons in the attractor network from the stimulus, compete successfully with the afferents generated by the recurrent activity inside the network, in a situation in which the both sets of synaptic efficacies are weak and approximately equal.

We simulate and analyze a two component network: one representing the stimulus, the other an ANN. We show that if stimuli are correlated with the receptive fields of neurons in the ANN, and are of sufficient contrast, the stimulus can provide the necessary information to the recurrent network to allow learning new stimuli, even in very disfavored situation of synaptic predominance in the recurrent part. Stimuli which are insufficiently correlated with the receptive fields, or are of insufficient contrast, are submerged by the recurrent activity.

1 Introduction

Attractors are playing an increasingly significant role in the interpretation of cortical activity, as well as in the description of cognitive phenomena\textsuperscript{1, 2}. The recent experiments of the Oxford group\textsuperscript{3} may even move the primary visual cortex from the simple, locally structured, feed forward network to the realm of intensive feed back attractor network (ANN). In fact, Dong and Hopfield\textsuperscript{4} have pointed out that if V1 is considered as a feedback dominated network, then allowing for adjustable synapses, a single step beyond the lateral geniculate nucleus (LGN), direction selectivity can be already accounted for.

Learning in attractor networks is a relatively new field. It is made complicated by the inter-dependence of two dynamical systems – neurons and synapses. It raises a whole host of new questions beyond those studied for ANN’s with fixed synapses. Those include the maintenance of learned synaptic values\textsuperscript{5}; the convergence of the double dynamics to stable states\textsuperscript{6} and many others. Treves and Rolls\textsuperscript{6} have raised another serious question: observing the CA3 region in hippocampus they note the following facts:

\footnotetext{\textsuperscript{1}On leave of absence from Racah Institute of Physics}
• Anatomically it is a perfect candidate for attractor dynamics, since the recurrent (lateral) connectivity is very intensive;

• The typical number of recurrent synapses per CA3 neuron is about four \(4\) times higher than the number of synapses per neuron coming from the main input source – the perifont path (PP);

• The synapses from recurrent contacts are typically closer to the soma than those coming from the perforant path;

• There is a secondary input route which seems to be carrying the same information as the PP and communicates it indirectly, via mossy fibers, to the same area CA3;

• Without the mossy fibers CA3 function, as a learning system, is significantly impaired.

Treves and Rolls[6] conclude that if the input through each synapse in the two sets of synapses (input and recurrent) is equal in amplitude, then the high ratio of recurrent to PP synapses \(4\) precludes the possibility of learning new items, because of the interference of the recurrent activity. This then requires the assistance of the mossy fibers, which carry similar information to the entering stimulus, but synapse more effectively on the neurons of CA3. It is further shown in ref. [6] that the relatively high recurrent connectivity does not impede retrieval of previously stored memories.

This description of the conditions in hippocampus is not uncontrovertial. Traub et al[7] review experimental evidence that synaptic efficacies in CA3 are very strong, essentially between all pairs of neurons, while the connectivity is very intense (each neuron is synapsing on thousands of other neurons). This would make the problem of learning and of the operation mode very different from what is familiar in the theory of ANNs. Yet, the arguments of Treves and Rolls are quite generic while the actual conditions in hippocampus or the availability of mossy fibers may not be. This raises a serious concern about the possibility of learning in attractor neural networks, be they biological or electronic, in conditions in which recurrent connectivity is more numerous than the afferent one and afferent synapses are not more effective than the recurrent ones.

We have therefore undertaken the study of a simple network which respects all the constraints listed above, namely the ratio of the synaptic numbers is the same and the average strength of excitatory synapses in the recurrent network and from the input to the recurrent network is equal. Our model network is composed of two parts: one like a retina (or an LGN), has the input represented on it as neural activities. It may also be considered as a cross section through the PP. The second is like CA3, with recurrent connections. The two networks are connected by feed forward connections, which are circular projections from the input network to
the recurrent network. This connection scheme creates receptive fields in the CA3 network. The recurrent network is then wired up to store a set of uncorrelated patterns, of active and passive neurons, à la Willshaw[8]. We have then proceeded to study the distribution of depolarization in the recurrent network, for different presentations of stimuli on the input network.

We have observed that the distribution of depolarization depended very strongly on the type of stimulus presented: for stimuli which are uncorrelated with the structure of the receptive fields, or for correlated stimuli which are not of high enough contrast, the conclusions of ref. [6] are confirmed. However, when such a correlation is present in the stimulus, and the stimulus is of sufficient contrast, the distribution has enough information to learn the new pattern, even if the CA3 was in one of the attractors previously learned. This we have seen in simulations, and as a consequence we have been able to deduce certain quantitative conditions on the relation between the correlations in the input and the coding rate of attractors in the recurrent network.

We have then proceeded (in Section 5) to investigate a simplified dynamics of the recurrent network, in order to ascertain that given a an inhibitory reaction which controls the overall spike rates in the recurrent network, the distribution of depolarizations created by an effective stimulus leads from a pre-existing reverberation in an attractor, in the absence of a stimulus, to a new attractor representing the stimulus quite faithfully. This attractor is not a fixed point, since the synaptic matrix is not symmetric, but the states it wanders over cover a very small space with a large overlap with the new stimulus.

Finally, we discuss the results in view of the fact that hippocampus does seem to require the mossy fibers. We also point out that the constraints on the input found here may be softened, if one relaxes some of the constraints imposed on the model network, such as equal fraction of active neurons in the input and in the recurrent network, or the equal spike frequencies in both networks.

2 The model network

We study a network composed of two groups of neurons, an input layer and a recurrent part.

The input layer: this part of the network is a two-dimensional grid of $N_I$ neurons and will be referred to as IN (input network). It mimics a cross section through the perforant path. These neurons present binary patterns of activity to the recurrent network. In this layer there are no interconnections and no dynamics. Stimuli are presented on this network by activating a fraction $f_I$ of the neurons with a fixed firing spike rate $\nu_I$; the remaining neurons have zero firing rate. We assume periodic boundary conditions for IN.

The recurrent network: abbreviated as RN, mimics CA3. This part is composed of $N_R$ neurons (for simplicity we take $N_R = N_I = N$). It stores a set of
p binary random patterns \( \eta^\mu (\mu = 1, \ldots, p) \). The stored patterns are \( N \)-bit words drawn independently, at each site \( i, (i - 1, \ldots, N) \), from the distribution

\[
P(\eta^\mu_i = 1) = f_R, \quad P(\eta^\mu_i = 0) = (1 - f_R)
\]

where \( f_R \) is the coding rate (the fraction of neurons active in an attractor) in this network. The active neurons in a recalled memory fire at a rate \( \nu_R \); the remaining neurons have zero firing rate.

Memory of these patterns is expressed on a randomly chosen subset of \( CN \) synapses out of the set of \( (N-1)N \) possible synapses. The \( (N-1-C_R)N \) remaining links are set to zero. The subset of links expressing the memory, \( \{ J^{R}_{ij} \} \), is set on top of a uniform excitatory background, according to the Willshaw prescription[8], i.e.

\[
J^{R}_{ij} = J^c + \begin{cases} J^R & \text{if both neurons } i \text{ and } j \text{ are active in at least one pattern} \\ 0 & \text{otherwise}. \end{cases}
\]

For this prescription the probability for any connection being \( J^0 + J^R \), in the subset of modifiable links, is \( c_R = 1 - \frac{N_R}{N} \) [8]. However, the average recurrent connectivity in RN is: \( C_R \).

It is the average excitatory synaptic strength

\[
c_R(J^0 + J^R) + (1 - c_R)J^0
\]

which will be set equal to the typical synaptic strength arriving from the input.

**Connections between IN and RN - receptive fields:** each of the \( N \) neurons of RN has a receptive field on IN centered on one of the IN neurons, and receives, on average, \( C_I \) afferent connections from neurons in IN. This average connectivity will satisfy the constraint

\[
C_I/C_R = K,
\]

where \( K \sim 0.25 \) in ref [6]. The procedure for drawing the synaptic matrix \( \{ J^I_{ij} \} \) \((i, j = 1, \ldots, N)\) is:

- We place the neurons of RN on a grid identical to that of IN. The position, \( \vec{x}_i \), of the neuron in IN at the center of the receptive field of neuron \( i \) in RN, is chosen to be the coordinate of neuron \( i \) in the identical grid of RN. Note that this does not imply any spatial structure for RN.

- If neuron \( j \) is at position \( \vec{x}_j \) in IN the link \( J^I_{ij} \), to neuron \( i \) in RN, will be present with probability

\[
P(J^I_{ij} \neq 0) = \frac{C_I}{2\pi \rho^2} \exp \left( -\frac{(\vec{x}_i - \vec{x}_j)^2}{2\rho^2} \right)
\]
where $\rho$ measures the size of the receptive field. The probability that these two neurons are not synaptically connected is:

$$P(J_{ij}^I = 0) = 1 - P(J_{ij}^I \neq 0)$$

The fact that the probability in the center of the receptive field cannot be greater than unity, produces a constraint on the minimal size of the receptive field $\rho$ for a given connectivity level $C_I$. Since $P(J_{ii}^I) = P_0 < 1$, we must have

$$\rho \geq \frac{C_I}{2\pi}$$

Moreover, the normalization of (4) ensures that summing the distribution over all $j$, i.e. over the input network, will produce, on average, $C_I$ links from IN for every neuron $i$ in RN.

- The existing links have uniform efficacies, i.e. $J_{ij}^I = J^I$.

In Fig. 1 we show the combined network, placed on two identical grids to facilitate the description of the receptive field of neurons in RN. Note that after the receptive fields have been established, the placing of the neurons in RN can be shuffled arbitrarily.
3 Distribution of the synaptic currents

3.1 The currents in RN

The contribution of the recurrent synapses to the current arriving in neuron \(i\) of RN can be expressed as

\[
h_i^R = \sum_j J_{ij}^R \nu_j^R
\]

where \(\nu_j^R\) is the activity of neuron \(j\) in RN. In what follows we will be taking bi-modal distributions for the neuronal activities in the attractors: the active neurons will have \(\nu_j^R = \nu_R\) and the others \(\nu_j^R = 0\). When RN is in one of the memorized attractors, in the absence of a stimulus, we have for example, \(\nu_i^R = \nu_R \eta_i^1\) for all neurons \(i\) in RN. This implies that all neurons active in this memory have maximal spike frequency \(\nu = \nu_R\), while all the others are quiescent. The resulting distribution of the afferent currents among the neurons of RN is shown in Fig. 2. What one observes is a bimodal distribution: neurons in the foreground of the retrieved memory (those for which \(\nu_i = \nu_R \eta_i^1 = \nu_R\)) have high afferent currents and subsequently will have elevated spike rates, while the neurons in the background (\(\nu_i = \nu_R \eta_i^1 = 0\)) receive low currents and would be essentially quiescent. This state of activity, in absence of a new stimulus, can be stabilized provided the neurons have an effective threshold lying in the gap between the two peaks. Such a threshold would require the introduction of inhibition, which we describe briefly in Section 5.

3.2 Necessary condition for acquisition of new patterns

We set aside the question of the maintenance of the attractors in the recurrent network, which goes beyond the considerations of ref. [6]. The question we pose is: what are the necessary effects of a new stimulus on the distribution of afferent currents in RN, to make acquisition possible? A minimal requirement would be that upon presentation of the new stimulus, a subset of neurons in RN, different from the retrieved memory \(\eta^1\), have particularly high combined afferent currents. In other words, a group of neurons, most of which were in the background, and had low afferent currents from the recurrent activity, join a potential smaller group of neurons in the foreground to have combined afferents (recurrent plus input) significantly higher than those of the neurons in the foreground in the absence of the stimulus.

This is a non trivial requirement, which we proceed to discuss in the next subsection. Here, we shall only add that if this situation can be realized, then a new threshold can separate the neurons marked in the presence of the stimulus from all the others. The stimulus can, in principle, be learned. The necessary
adjustment of the threshold is again the task of the associated inhibitory network. See e.g. Section 5.

3.3 Structure and effect of stimuli

The incoming current, due to a stimulus presented on IN, to a neuron $i$ of the recurrent network is given by

$$h_i^l = \sum_j J_{ij}^l \nu_j^l$$

where $\nu_j^l$ is the activity of neuron $j$ in IN. The distribution of these incoming currents depends on the stimulus presented.

In the following we will consider stimuli activating a circular spot of neurons on the grid, of radius $r$. Inside the circle the fraction of active neurons $p_s$ is higher than the average fraction of active input neurons $f_I$. Outside the circle the fraction of active neurons $p_b$ is lower than $f_I$, in order to keep the global activity fixed. Neurons which are activated by the stimulus have equal frequencies: $\nu_j^f = \nu_I$. An example is shown in Fig. 3. More precisely, the patterns in the simulations are generated by the following procedure: first the center of the spot is drawn randomly, then the radius $r$ of the spot is determined by the condition that it covers about $f_I N$ neurons. The size of the spot is thus determined by the relation $\pi r^2 \sim f_I N$. The probability that a neuron is active is $p_s$ inside the spot and $p_b$ outside, with
\[ p = p_e \]

spot

\[ p = p_s = p_e + \lambda \quad \text{RF} \]

\[ r \quad \rho \]

Figure 3: Stimulus and receptive field represented on IN: a spot of activated neurons (circle of radius \( r \)) and the receptive field of some neuron in RN (schematized by the circle RF of radius \( \rho \)). The probability \( p \) of a neuron being active is indicated inside and outside the circle.

\[ p_s = f_I + \lambda(1 - f_I), \quad p_e = (1 - \lambda)f_I \quad (8) \]

and \( 0 \leq \lambda \leq 1 \). Note that for every value of \( \lambda \) the global activity is the same, since the radius of the spot is kept fixed. For \( \lambda = 0 \) the pattern has no spatial structure: activity is distributed uniformly. For \( \lambda = 1 \) only neurons inside the spot are active while all others are quiet.

We point out that our way of constructing the stimulus implies that for fixed activity rate, \( f_I \), in IN the stimulus has a fixed radius. The stimulus can be varied only by varying \( \lambda \), i.e. the contrast. This suppressed degree of freedom is compensated by variation of the size of the receptive fields (see below). The results depend on the ratio of the two radii only.

We consider a particular neuron of RN, and see how its incoming current changes depending on the position of the presented stimulus. We take \( \rho = 6 \), which, with \( C_I = 200 \), makes the probability for the presence of the link connecting the neuron in RN to the center of its receptive field close to unity. We also assume, as in [6], that the incoming and the recurrent synapses have similar strengths. The size of the spot, with \( N = 1600 \) and \( f_I = 0.1 \), is \( r \sim 7 \) grid units. Fig. 4 shows the variation of the incoming current to a given RN neuron with the distance \( d \) (in grid units) from the center of the spot to the center of the receptive field of the neuron, for different values of \( \lambda \). The spot is translated on the grid along a horizontal line which includes the center of the receptive field. This is a bona fide tuning curve.

The distribution of both incoming currents and total (recurrent plus input) currents is considered in the two extreme cases \( \lambda = 0 \) (pattern without spatial
structure) and $\lambda = 1$ (full spot).

For $\lambda = 0$, the integrated input current in any neuron of RN is the sum of random uncorrelated variables. Thus when the number of incoming channels is large, the current from the input becomes a random gaussian variable with average $\langle h^I \rangle = J^I \nu_I f_I C_I$ and variance $\langle (\Delta h^I)^2 \rangle = f_I C_I (J^I \nu_I)^2 (1 - f_I P_c / 2)$, for any value of the size of the receptive fields $\rho$ (see e.g. appendix). The distribution of both incoming and total (recurrent plus incoming) currents, when RN is in one of its stored attractors, is shown in Fig. 5. The light curve (on the left) is the distribution of the afferent current due to the structureless stimulus. The bold curve is the effect the stimulus has when combined with the currents of the attractor, Fig. 2. In this case the effect of the recurrent connections eliminates all the information contained in IN, since the new bimodal distribution is just the shifted old one, dominated by the stored pattern.

For $\lambda = 1$, we consider the two cases $\rho > r$ and $\rho < r$. Note that because the average connectivity $C_I$ is fixed, the probability of a contact decreases with increasing $\rho$. Thus, if the size of the receptive field is small the probability of contact in its center will be high. Fig. 6 shows the distribution of the currents in the case of a small spot $r = 7 < \rho = 15$. Fig. 7 shows the distribution when the spot is large compared to the receptive field $r = 7 > \rho = 4$. (Note that with these parameters the condition (5) is not satisfied and $P_0$ would be larger than 1. The parameters were chosen for convenience in the simulation and the Gaussian, Eq. 4, was slightly modified. It was truncated to one where it would have become larger.
Figure 5: Distribution of incoming and total currents in an RN neuron (relative to the uniform excitatory background), for $\lambda = 0$ and the same parameters as in Figs. 2, 4. The light curve is the incoming current distribution. The bold curve is the total current distribution.

and care was taken so that the normalization ensure that the average connectivity remains equal to $C_I$.

In the case $\rho > r$, Fig. 6, the distribution of both incoming and total currents is similar to the uniform case. The reason is that the size of the spot is small compared to the size of the receptive fields, thus the difference of incoming currents between a neuron close to the spot and a neuron far from it will be small, compared to the difference of currents between foreground and background neurons in the attractor. We recall that the distance between a neuron and the spot is the distance between the center of the receptive field of the neuron and the center of the spot in IN.

In the opposite case (Fig. 7), $\rho < r$, the few neurons in RN that are close to the spot receive their input only from the active neurons of the spot and thus receive a much higher current than the neurons that are far from it, creating the tail of both distributions. In this case the neurons in RN that receive the highest total currents, i.e. those that have their currents in the tail of the distribution, are not anymore the neurons that were active before the presentation of the stimulus. The latter neurons have their currents in the peak at $h = 160$. The ones in the high tail are those that ‘see’ the most active region in IN. In this case the information contained in the input is preserved, making the learning of the new pattern possible.
Figure 6: Distribution of incoming and total currents in a RN neuron, for $\lambda = 1$, $\rho = 15 > r$. Other parameters as in Figs. 2, 4. Light curve: incoming current distribution. Bold curve: total current distribution.

Figure 7: Distribution of incoming and total currents in an RN neuron, for $\lambda = 1$, $\rho = 4 < r$. Other parameters as in Figs. 2, 4. Light curve: incoming current distribution. Bold curve: total current distribution.
4 Quantitative estimates

To arrive at a quantitative estimate of the constraints, we have to evaluate the currents coming from IN to every neuron of RN. We consider a circular stimulus of radius $r$, centered at the origin. The stimulus is defined by prescribing the probability that it activates a neuron in IN (i.e., raises its spike rate). For a neuron of IN with coordinates $x, y$, relative to the center of the spot ($x^2 + y^2 = s^2$), we take this probability to be $p_b + \lambda$ if $s < r$ (inside the spot) and $p_b$ if $s > r$ (outside the spot). The mean incoming current for a neuron of RN, with coordinates $(x, y)$, at distance $s$ from the origin, is

$$ \langle h^I(s) \rangle = C_IJ^I \nu_I([p_b + \lambda g(\alpha, s)]$$

where

$$g(\alpha, s) = \int \int_{X^2 + Y^2 < r^2} dXdY \frac{1}{2\pi \rho^2} \exp \left( -\frac{(X - x)^2 + (Y - y)^2}{2\rho^2} \right)$$

is the integral of the normal distribution over the circular spot, and $\alpha = r/\rho$. See e.g. Appendix.

The variance of the current for a neuron at distance $s$ from the origin is

$$\langle (\Delta h^I(s))^2 \rangle = (J^I \nu_I)^2 C_I[p_b + \lambda \psi(s)]$$

where $\psi(s) < 1$ for all $s$, as is shown in the Appendix. Consider the extreme case $r \gg \rho$, i.e., the large $\alpha$ limit. In this case the function $g$ is well approximated as follows: for $0 < s < r$

$$1 - g(\alpha, s) \sim \frac{r}{\sqrt{2\pi \alpha(r - s)}} \exp \left( -\frac{\alpha^2 (r - s)^2}{2r^2} \right)$$

and for $s > r$

$$g(\alpha, s) \sim \frac{r}{\sqrt{2\pi \alpha(s - r)}} \exp \left( -\frac{\alpha^2 (r - s)^2}{2r^2} \right).$$

So for very large $\alpha$,

$$g(\alpha, s) = \begin{cases} 1 & \text{if } s < r \\ 0 & \text{if } s > r. \end{cases}$$

An RN neuron with $s < r$, whose receptive field overlaps strongly with the stimulus, receives from IN a mean incoming current

$$\langle h^I(s) \rangle = C_IJ^I \nu_I(p_b + \lambda)$$

since the mean activity in its receptive field is $(p_b + \lambda)\nu_I$. The other neurons (those for which $s > r$) receive a mean current

$$\langle h^I(s) \rangle = p_b C_IJ^I \nu_I.$$
The difference of the mean input currents arriving at neurons that ‘see’ the stimulus and those that do not is, for large $\alpha$:

$$\Delta \langle h^I \rangle = \lambda C_I J^I \nu_I$$

But the input currents fluctuate. For an RN neuron with $s < r$, the RMS of the fluctuations is

$$\sqrt{\langle (\Delta h^I)^2 \rangle} = J^I \nu_I \sqrt{C_I p_w},$$

for a neuron with $s > r$ it is

$$\sqrt{\langle (\Delta h^I)^2 \rangle} = J^I \nu_I \sqrt{C_I (p_b + \lambda)}.$$

Consider first the case when $C_I$ is very large. In this case the fluctuations can be ignored, because the difference between the mean currents is of order $C_I$, while the fluctuations are of order $\sqrt{C_I}$. The average recurrent afferent current for neurons in the foreground and in the background, respectively, is:

$$C_R f_R (J^0 + J^R) \nu_R, \quad C_R f_R (J^0 + J^R \Delta) \nu_R,$$

where

$$\Delta = \frac{c_R - f^2}{1 - f^2}$$

for the synaptic prescription (2). Thus the difference between the average currents of neurons in the background and in the foreground is $f_R C_R J^R \nu_R (1 - \Delta)$.

The condition for learnability is that for a sufficient number of background neurons the afferent current be driven above the average current arriving at a foreground neuron whose receptive field is away from the spot illuminated by the stimulus to be learned. This condition is:

$$\lambda C_I J^I \nu_I > f_R C_R (1 - \Delta) J^R \nu_R.$$

It can be expressed as a condition on the ‘contrast’ of the spot. Substituting the constraint (3), we find that the stimulus is learnable if:

$$\lambda > \lambda_0 = \frac{f_R (1 - \Delta) J^R \nu_R}{K J^I \nu_I}. \quad (10)$$

If this condition is satisfied, neurons that have the highest contribution in the distribution of the currents will be the ones that receive their inputs from the high activity spot of the IN layer, regardless of whether they were in the foreground or in the background before the presentation of the stimulus. Taking $K = 0.25$ as in [6], $\nu_R = \nu_I$, and $J^R = J^I$, the condition becomes: $\lambda > 4 f_R (1 - \Delta)$. In other words, the higher the advantage of recurrence over input connectivity (i.e. the lower $K$), the higher must the contrast of the spot be for the stimulus to be learned. Not a very great surprise.
Note that for a large but finite \( \alpha \) the distribution of the input currents will not be two-valued anymore, but neurons close to the boundary of the spot will receive currents with intermediate values between the high and low currents. Still for \( \alpha \) large enough a significant fraction of neurons inside the spot will receive a current very close to \( C_I J^I \nu_I(p_b + \lambda) \).

Now we consider the fluctuations in the input currents, for large \( \alpha \). These fluctuations are of order \( \sqrt{C_I p_b} \) outside and \( \sqrt{C_I (p_b + \lambda)} \) inside the spot. For these fluctuations not to affect learnability, they should not close the gap between the high and the low currents. The resulting condition is:

\[
\sqrt{C_I (\lambda - \lambda_0)} > \sqrt{p_b} + \sqrt{p_b + \lambda},
\]

If the contrast \( \lambda \) is high compared to the coding rate \( f_R \), as implied by the condition \( \lambda > 4 f_R (1 - \Delta) \), and if \( f_R = f_I \), then \( \lambda \gg p_b \) since Eq. (8) implies \( p_b < f_I \). Thus the condition reduces to

\[
\lambda > \lambda_0 \left(1 + \frac{1 + \sqrt{1 + 4 \lambda_0 C_I}}{2 \lambda_0 C_I}\right).
\]

Moreover if \( C_I \lambda_0 \gg 1 \), which seems to be a realistic condition (for example in the CA3 of the rat \( C_I \sim 3000 \) [6], and probably \( \lambda_0 > 0.01 \)), the condition simplifies to

\[
\lambda > \lambda_0 \left(1 + \frac{1}{\sqrt{\lambda_0 C_I}}\right).
\]

It is quite interesting that even when the fluctuations are of the order of the difference of the currents, or when the condition \( \rho \gg \rho \) does not hold (as in all the simulations where \( \alpha \) is of order one), all is not lost. In that case, not all the neurons ‘seeing’ the high-activity spot will receive a higher current than the foreground neurons, upon presentation of the stimulus. However, if condition (9) holds, a significant fraction of them will receive a higher current than the foreground neurons, and the attractor in RN will be destabilized. In this case, the difference between the recurrent currents in background and foreground neurons will decrease in time. If the stimulus persists, the number of active neurons seeing the spot will increase, till a completely new configuration, highly correlated with the stimulus, is stabilized. We show in the next section that this indeed happens, for a very simple type of dynamics in the recurrent network.

Note that the effective threshold of the neurons in the recurrent network has to increase when a correlated pattern is presented. Otherwise, both the foreground neurons and the ones that see the high-activity regions in IN will be active. If the threshold increases, the network will separate these two types of neurons and may, in principle, be able to learn the ‘new’ pattern. For example, in the case discussed earlier the effective threshold has to increase from an initial value lower than \( \nu_{0, R} C_R (J^0 + J^R) \nu_R \) to some value between \( \nu_{0, R} C_R (J^0 + J^R) \nu_R + p_c C_I J^I \nu_I \) and \( \nu_{0, R} C_R (J^0 + J^R \Delta) \nu_R + p_c C_I J^I \nu_I \).
5 Dynamics of the recurrent network in presence of stimulus

We describe the simulation of a very simple dynamics in the recurrent layer, following the presentation of a stimulus. It shows that if the input pattern is selective enough, a new configuration, independent of any previously learned patterns, is stabilized in RN. In contrast, if the pattern is weakly correlated with the receptive fields, the state of the recurrent network is not modified by the presentation of the stimulus.

In the above we have considered a recurrent network made only of excitatory neurons, or only excitatory synapses. Such a network, when iterated will be extremely unstable. Usually, one introduces an activity dependent threshold, to separate the foreground from the background neurons. In a real network the activity dependent threshold is likely to be a reactive inhibition (see e.g. [9, 10]). We adopt this attitude also in the model considered in this section.

To the network defined in section 2 we attach a network of $N_{inh}$ inhibitory neurons connected to all the neurons in RN. The synapses connecting the excitatory neurons to the inhibitory ones are taken to be all equal, i.e.

$$J_{ij}^{ei} = J_{i}^{e} \quad i = 1, \ldots, N_{inh} \quad j = 1, \ldots, N$$

also the inhibitory-excitatory synapses are all equal:

$$J_{ij}^{ei} = J_{i}^{e} \quad i = 1, \ldots, N \quad j = 1, \ldots, N_{inh}$$

as well as the inhibitory-inhibitory ones:

$$J_{ij}^{ii} = J_{i}^{ii} \quad i, j = 1, \ldots, N_{inh}.$$ 

The current arriving at an excitatory neuron $i$, in RN, at time $t$ is

$$k_{i}^{exc}(t) = \sum_{j \in IN} J_{ij}^{e} v_{j}^{I}(t) + \sum_{j \in RN} J_{ij}^{R} v_{j}^{R}(t) - J_{i}^{e} \sum_{j=1}^{N_{inh}} v_{j}^{inh}(t).$$

Note that the last term in this equation, representing the contribution of the inhibitory neurons to the current, is effectively a time (activity)-dependent contribution to the threshold. For the inhibitory neuron $i$ the current at time $t$ is

$$k_{i}^{inh}(t) = J_{ij}^{e} \sum_{j \in RN} v_{j}^{R}(t) - J_{i}^{ii} \sum_{j=1}^{N_{inh}} v_{j}^{inh}(t).$$

Thus all neurons in the inhibitory network receive the same excitatory current and every excitatory neuron receives the same inhibitory afferent. Consequently, we can consider only one inhibitory neuron. Its activity $a_{inh}(t)$ is determined at time $t$ by
\[ a_{\text{inh}}(t) - \phi(g_{\text{ie}}a_R(t) - g_{\text{ia}}a_{\text{inh}}(t)) \]  

where \( g_{\text{ie}} = J^{\text{ie}}N \), \( g_{\text{ia}} = J^{\text{ia}}N_{\text{inh}} \), and \( a_R(t) \) is the mean activity in the recurrent, excitatory network:

\[ a_R(t) = \frac{1}{N} \sum_{i \in \text{RN}} v^R_i(t). \]

The current in an excitatory neuron \( i \) is

\[ h^\text{exc}_i(t) = \sum_{j \in \text{IN}} J^I_{ij}v^I_j(t) + \sum_{j \in \text{RN}} J^R_{ij}v^R_j(t) - g_{ei}a_{\text{inh}}(t) \]

where \( g_{ei} = J^{ei}N_{\text{inh}} \). This network is equivalent to RN plus one effective inhibitory neuron whose activity is determined by Eq. (11), coupled to every neuron in RN with synaptic strength \( g_{ei} \). Again, this inhibitory neuron acts effectively as a time (activity)-dependent threshold.

The RN neurons have a fixed, uniform threshold \( T \); their spike rate at time \( t + 1 \) is determined by

\[ v^R_i(t) = \begin{cases} v_R & \text{if } h_i(t) > T \\ 0 & \text{otherwise} \end{cases}, \]

Note that the dynamics of these neurons is effectively that of binary units. This is done in order to avoid the additional complications of rate dynamics, see e.g. ref. [11]. We permit ourselves this simplification because our purpose here is not to describe a network of realistic neurons, but merely to produce an indication that the classifications of neurons by the afferents is not an artifact of the initial presentation of the stimulus. The value of \( v_R \) is that value of the frequency which we would expect the network to find as its stationary one in an attractor with a bi-modal distribution. The activity of the inhibitory neuron is determined at each time step by equation (11).

Before the presentation of the stimulus, RN is in one of the stored attractors, for example \( \eta^1 \). We present a stimulus of the type described in Section 3. (i.e. circular spots of radius \( r \) and contrast parameter \( \lambda \)) and observe the temporal evolution of the following parameters:

- the mean activity, \( a_R(t) \), in RN and the activity of the inhibitory neuron \( a_{\text{inh}}(t) \);
- the mean activity of the foreground and background neurons

\[ m_+(t) = \frac{1}{f_RN} \sum_{i=1}^N \eta^1_i v^R_i(t) \]
Figure 8: Temporal evolution of activity in RN: overlaps with memorized pattern, background and stimulus, respectively: $m_+ (\bigodot), m_c (\bigoplus)$ and $m_s (\square)$, averaged over 100 stimuli. $N = 900$ neurons ($30 \times 30$ grid), both for IN and for RN. The recurrent and input connectivities are, respectively: $C_R = 400, C_I = 100$. Coding rates: $f_R = f_I = f = 0.1$; synaptic efficacies: $J^0 = 0.9, J^I = J^R = 1$; couplings to inhibitory part: $g_{ee} = 1.7, g_{ii} = 0, g_{ei} = 300$; excitatory threshold: $T = 40$; activity rates in IN and RN: $\nu_I = \nu_R = 1$; receptive fields: $\rho = 3.5$ grid units. The stimulus is a circular spot with $r \sim 5$ (determined approximately by $N f_I$), $\lambda = 0.1$.

$$m_0(t) = \frac{1}{(1 - f_R)N} \sum_{i=1}^{N} (1 - \eta_i^1)\nu_i^R(t)$$

- the fraction $m_s(t)$ of active neurons whose center of receptive field is inside the spot.

We have simulated the dynamics of the combined network, starting always from one of the 10 patterns memorized in the 900 neurons of RN and averaging the different overlaps, quantifying the dynamical evolution, over 100 runs. The fraction of synapses modified by the Willshaw prescription is: $c_R = 0.1$. The response of the inhibitory neurons to the mean rates in RN was chosen, for the sake of the illustration, to be:

$$\phi(x) = \tanh(x - \theta)$$

with $\theta = 0.07$.

The results are presented in Figs. 8, 9, in which the values of all the parameters are specified. The behavior of this network depends strongly on the presented stim-
Figure 9: Temporal evolution of the same network as in Fig. 8, with \( \lambda = 1 \).

ulius. We observe two types of behaviour depending on the brightness parameter \( \lambda \):

When \( \lambda \) is weak \((\lambda < 0.5)\) (Fig. 8, \( \lambda = 0.1 \)) the state of the recurrent network is weakly modified by the stimulus. Mainly it wanders in the neighborhood of the memorized attractor, as witnessed by the high average value about which \( m_+ \) (\( \bigtriangleup \)) fluctuates. Sometimes a few neurons of the background are active or oscillate: the activity \( a_R(t) \) decreases slowly from 0.1 to 0.07, but we have for all time steps \( m_+ \sim a_R/t, m_0 \ll 0.1, m_s \sim 0.1 \).

On the other hand, if \( \lambda > 0.5 \) (Fig. 9, \( \lambda = 1.0 \)) the network evolves from the initial state, a stored pattern, towards a configuration highly correlated with the input spot: following a few time steps, we have \( m_+ \sim m_c \sim 0.1 \) (\( \bigtriangleup \)), which is a decrease towards the random overlap between the stored patterns. On the other hand, 0.1 of the background neurons \( m_c, +'s \) are activated, which corresponds to the neurons driven by the new stimulus. This is witnessed by the rapid increase of \( m_s \) (\( \Box \)). The asymptotic value of \( m_s \) increases with \( \lambda \) from \( m_s \sim 0.4 \) when \( \lambda = 0.6 \) to \( m_s \sim 0.8 \) when \( \lambda = 1 \) (Fig. 9). The network almost never goes to a fixed point; however after a few time steps only a small number of neurons (about 10 out of 900) oscillate between active and inactive states. The activity \( a_R(t) \) increases initially, but then decreases to a value close to 0.1.

We show the time evolution of the parameters \( m_+, m_0 \) and \( m_s \), averaged over 100 stimuli, in the two cases \( \lambda = 0.1 \) (Fig. 8) and \( \lambda = 1 \) (Fig. 9). The error bars represent the fluctuations between different stimuli.
6 Discussion

The simple model described in this study does not resolve any of the puzzles of hippocampus. In the context of a description of hippocampal connectivity as in ref. [6], it puts the focus on the structure of the information arriving via the perforant path. If, in fact, hippocampus cannot learn without the mossy fibers[12], then our study would lead one to conclude that information arriving along the perforant path is not well organized relative to the receptive fields of neurons in CA3, provided such receptive fields exist in CA3 in the first place. Both questions seem to be empirically accessible. If the connectivity in CA3 accords with the description of ref. [7], this study is totally irrelevant for hippocampus.

On the other hand, this study concludes that whatever the difficulties of learning in hippocampus may be, they do not represent a universal limitation on learning in ANN’s. Usually, cortical communication is organized with receptive fields. Coding rates in attractors may be rather low, so the constraint on the level of contrast, \( \lambda \) in the stimulus is not very severe. This constraint becomes even less binding when one allows:

- Higher coding rate on the input line than in the ANN;
- Higher spike frequencies in neurons carrying the stimulus, i.e. the analogs of the perforant path;
- Higher rates on the neurons of the recurrent network, due to the dynamics ensuing the new distribution of afferent currents.

Our choice of the Willshaw synaptic matrix may appear somewhat particular. We believe that though it is unlikely to represent synaptic organization in detail, it does contain some very generic features about representing learned information in the purely excitatory part of the synaptic arrangement. This is also connected to our assumption that both in the input channel and in the ANN significant activity is distributed in a bi-modal fashion. Such bi-modality is not observed in cortex[13]. Yet, this does not invalidate our considerations, which can be expressed in terms of distributions of rates and non-uniform synaptic values.

Moreover, one may object that the random patterns composing the recurrent synaptic matrix are not representative of the patterns that are considered for learning and which would have been learned in the past. A more complete treatment would have the recurrent network learn what is learnable from the geometrically organized stimuli and then test the learnability of consecutive stimuli. We have avoided this task for several reasons. First, because it would have taken us way beyond the scope of the problem posed. Second, it would have required a commitment to a given learning dynamics. Third, we felt that to test the learnability of an input stimulus which is uncorrelated with the previously learned stimuli it
Figure 10: Circular spot in IN (circle of radius \( r \)), with one RN neuron placed for convenience at the location of the center of its receptive field in IN. The center of the receptive field has coordinates \( \vec{x}_i = (x, y) \) relative to the center of the spot. \( (s^2 = x^2 + y^2) \). The receptive field of the RN neuron is the circle of radius \( \rho \) (case \( r \gg \rho \)). Left: \( s < r \). Right: \( s > r \).

may be sufficient to consider random patterns in RN, provided the coding level is appropriate. Yet the fuller program is worth investigating.

Finally, the dynamics is very schematized, ignoring all the intricacies of inhibitory response times and transmission delays (see e.g. ref. [10]). Nor did we try to explore the parameter space of interactions. Our aim was limited to showing that the recurrent dynamics, following the presentation of a new, adequate stimulus, preserves the information presented for learning.

Acknowledgements

We are indebted to Stefano Fusi and Dr. Roger Traub for many helpful discussions and to Alessandro Treves and an anonymous referee for very helpful comments on the manuscript. N.B. gratefully acknowledges a fellowship of the EEC.

Appendix

Here we calculate the two first moments of the distribution of the input currents, for an RN neuron \( i \) with coordinates \( \vec{x}_i = (x, y) \), relative to the center of the spot, \( (s^2 = x^2 + y^2) \). Fig. 10 represents the circular spot together with RN neuron \( i \) and its receptive field, placed for convenience on the same plane, in the two cases \( s < r \) and \( s > r \).
The input current is
\[ h_i^I = \sum_{j \in \mathbb{IN}} J_{ij} \nu_j \]
The \( J \)'s and \( \nu \)'s are independent from each other and independent at each site. Hence, the mean of the input current at site \( i \) is given by
\[ \langle h_i^I \rangle = \sum_{j \in \mathbb{IN}} \langle J_{ij} \rangle \langle \nu_j \rangle. \] (13)

Using \( \langle \nu_j \rangle = \nu I p_i \) if \( j \) is outside the spot and \( \langle \nu_j \rangle = \nu I (p_i + \lambda) \) if \( j \) is inside the spot, we obtain
\[ \langle h_i^I \rangle = J^I \nu I p_i C_I + \nu I \lambda \sum_{j \in \text{spot}} \langle J_{ij} \rangle. \] (14)

For the neurons inside the spot, we have
\[ \sum_{j \in \text{spot}} \langle J_{ij} \rangle = J^I C_I \sum_{j \in \text{spot}} \frac{1}{2\pi \rho^2} \exp \left( -\frac{1}{2\rho^2} (\bar{x}_i - \bar{x}_j)^2 \right) \] (15)
where \( \bar{x}_j \) is the position of \( \mathbb{IN} \) neuron \( j \). When the number of neurons inside the spot is large, we can approximate the sum by the integral \( g(\alpha, s) \) of the normal distribution over the circular spot:
\[ g(\alpha, s) = \int \int_{X^2 + Y^2 < \alpha^2} \frac{n(X, Y)}{2\pi \rho^2} \exp \left( -\frac{1}{2\rho^2} ((X - x)^2 + (Y - y)^2) \right) \] (16)
where \( n(X, Y) \) is the surface density of neurons at \( (X, Y) \) and \( \alpha = r/\rho \). The density is taken to be uniform on the plane and the unit of distance is chosen such that \( n = 1 \). Thus, when (16) and (15) are substituted in Eq. (14), we obtain
\[ \langle h^I(s) \rangle = J^I \nu I C_I \rho I + \lambda g(\alpha, s). \]

For the calculation of the second moment we use
\[ \langle h_i^I \rangle^2 = \sum_{j, k \in \mathbb{IN}} J_{ij} J_{ik} \nu_j \nu_k \]
and the variance is
\[ \langle (\Delta h_i^I)^2 \rangle = \sum_{j \in \mathbb{IN}} \left[ \langle J_{ij}^2 \rangle \langle \nu_j^2 \rangle - \langle J_{ij} \rangle^2 \langle \nu_j \rangle^2 \right]. \] (17)
The sums are separated over neurons outside and inside the spot. Outside the spot \( \langle \nu_j^2 \rangle = \nu I \rho_i I \), and inside \( \langle \nu_j^2 \rangle = (p_i + \lambda) \nu_j I \). Furthermore,
\[ \sum_{j \in \text{spot}} \langle J_{ij}^2 \rangle = (J^I)^2 \rho \sum_{j \in \text{spot}} \exp \left( -\frac{1}{2\rho^2} (\bar{x}_i - \bar{x}_j)^2 \right) \]

\[ \sum_{j, k \in \mathbb{IN}} J_{ij} J_{ik} \nu_j \nu_k \]
and the variance is
\[ \langle (\Delta h_i^I)^2 \rangle = \sum_{j \in \mathbb{IN}} \left[ \langle J_{ij}^2 \rangle \langle \nu_j^2 \rangle - \langle J_{ij} \rangle^2 \langle \nu_j \rangle^2 \right]. \] (17)
The sums are separated over neurons outside and inside the spot. Outside the spot \( \langle \nu_j^2 \rangle = \nu I \rho_i I \), and inside \( \langle \nu_j^2 \rangle = (p_i + \lambda) \nu_j I \). Furthermore,
\[ \sum_{j \in \text{spot}} \langle J_{ij}^2 \rangle = (J^I)^2 \rho \sum_{j \in \text{spot}} \exp \left( -\frac{1}{2\rho^2} (\bar{x}_i - \bar{x}_j)^2 \right) \]
\[
\sum_{j \in \text{spot}} \langle J_{ij} \rangle^2 = (J^I_0)^2 P_0 \sum_{j \in \text{spot}} \exp \left( -\frac{1}{\rho^2} (\bar{x}_i - \bar{x}_j)^2 \right)
\]

where \( P_0 \) is the probability, Eq. 4, at the center of the receptive field, i.e. \( P_0 = C_I/(2\pi \rho^2) \). The sums are replaced by the integrals and we get

\[
\sum_{j \in \text{spot}} \langle J_{ij} \rangle^2 = (J^I_0)^2 C_I g(\alpha, s)
\]

and

\[
\sum_{j \in \text{spot}} \langle J_{ij} \rangle^2 = \frac{1}{2} (J^I_0)^2 C_I P_0 g \left( \frac{\alpha}{\sqrt{2}}, s \right).
\]

Substituting these equations in (17) we find

\[
\langle (\Delta h^I(s))^2 \rangle = (J^I_{1l})^2 C_I \left[ p_l + \lambda g(\alpha, s) \right] \psi(\alpha, s)
\]

where

\[
\psi(\alpha, s) = 1 - \frac{P_0 \left[ p_l^2 + \lambda(\lambda + 2p_l) g \left( \frac{\alpha}{\sqrt{2}}, s \right) \right]}{2[p_l + \lambda g(\alpha, s)]}
\]

Thus \( \psi(\alpha, s) < 1 \) for every \( s \) and \( \alpha \). Furthermore, \( \psi(\alpha, s) > 1 - P_0(p_l + \lambda)/2 \) since for every \( s \) \( g(\alpha, s) \) is a decreasing function of \( \alpha \).

References


[6] Treves A and Rolls E T 1992 Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network, Hippocampus 2 625

22


