Network: Comput. Neural Syst. **9** (1998) 123–152. Printed in the UK PII: S0954-898X(98)82605-5

Slow stochastic Hebbian learning of classes of stimuli in a recurrent neural network

Nicolas Brunel†, Francesco Carusi‡ and Stefano Fusi‡§‖

† LPS¶, Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France
‡ INFN, Sezione dell'Università di Roma 'La Sapienza', Pza Aldo Moro 2, 00185, Rome, Italy
§ Racah Institute of Physics, Hebrew University of Jerusalem

Received 14 March 1997

Abstract. We study unsupervised Hebbian learning in a recurrent network in which synapses have a finite number of stable states. Stimuli received by the network are drawn at random at each presentation from a set of classes. Each class is defined as a cluster in stimulus space, centred on the class prototype. The presentation protocol is chosen to mimic the protocols of visual memory experiments in which a set of stimuli is presented repeatedly in a random way. The statistics of the input stream may be stationary, or changing. Each stimulus induces, in a stochastic way, transitions between stable synaptic states. Learning dynamics is studied analytically in the *slow learning* limit, in which a given stimulus has to be presented many times before it is memorized, i.e. before synaptic modifications enable a pattern of activity correlated with the stimulus to become an attractor of the recurrent network. We show that in this limit the synaptic matrix becomes more correlated with the class prototypes than with any of the instances of the class. We also show that the number of classes that can be learned increases sharply when the coding level decreases, and determine the speeds of learning and forgetting of classes in the case of changes in the statistics of the input stream.

1. Introduction

It is widely believed that synaptic plasticity is the basic phenomenon underlying learning and memory. There is experimental evidence that neuronal activity can affect synaptic strength, through both long-term potentiation (LTP, e.g. Bliss and Collingridge 1993) and long-term depression (LTD, e.g. Artola and Singer 1993, Christie *et al* 1994). A large number of learning 'rules', specifying how activity and training experience change synaptic efficacies, has been proposed (Hebb 1949, Sejnowski 1977, Bienenstock *et al* 1982); such learning rules have been essential for the construction of most models of associative memory (Hopfield 1982, Amit 1989). In such models, the presentation of a stimulus to a recurrent network provokes modifications in the efficacy of recurrent collaterals. These modifications enable the pattern of activity evoked by the stimulus to sustain itself after the removal of the stimulus; this pattern becomes an attractor of the system. Such 'active memory states' have been observed in several areas of association cortex of monkeys performing delay memory tasks (Miyashita 1993, Fuster 1995). Recently, more realistic associative memory models have been developed (Amit *et al* 1994, Amit and Brunel 1997), allowing for quantitative comparisons with available experimental data.

‖ Part of PhD Thesis at Hebrew University.
¶ Laboratory associated with CNRS, Paris 6 and Paris 7 Universities.

0954-898X/98/010123+30$19.50

Since the structure of the synaptic efficacies is an essential characteristic of these models, the development of a realistic synaptic dynamics is particularly important. Most models of associative memory have left aside this issue, usually taking for granted a 'covariance-type' synaptic matrix (Hopfield 1982, Tsodyks and Feigel'man 1988), with the underlying assumption that any single synapse is able to preserve an arbitrarily high number of stable states over long timescales. A notable exception is the Willshaw model (Willshaw *et al* 1969), in which each synaptic efficacy has only two states, but again this model leaves aside the question of the dynamics. Shinomoto (1987) has introduced an on-line learning dynamics, but in this model the synaptic efficacies must have a long time constant and the stimuli must be presented continuously, otherwise the network returns to the initial (before learning) state. Nadal *et al* (1986) and Parisi (1986) showed that modifications of the Hopfield model to prevent memory blackout during sequential learning of patterns could make the network exhibit the 'palimpsest' property (old stimuli are automatically forgotten to make room for the most recent ones). In Nadal *et al* (1986) at each presentation the new synaptic efficacy is the sum of the old value multiplied by a decay factor plus the contribution of the new stimulus. In Parisi (1986) the 'palimpsest' property was obtained by bounding the synaptic efficacies.

More recently, the idea that synaptic efficacies have only a limited number of stable states and that a stimulus arriving at the network provokes transitions between these states began to develop as an alternate, more realistic, description of the learning process. Tsodyks (1990) showed that a drastic reduction in the memory capacity is inherent to such a learning process, as compared with the usual 'covariance' matrix. In this context, Wong *et al* (1991) studied the dependence of short-term memory behaviour on the initial synaptic distribution. In their study stimuli were learned in one shot and with certain choices of parameters the network exhibited primacy or recency phenomena. The idea of the learning process as a random walk was also considered, in a more general context, by Heskes and Kappen (1991).

Amit and Fusi (1992, 1994) have studied a learning process in which each synapse has only a finite number of stable states on long timescales, and neural activities induce stochastic transitions between these states, in the situation in which stimuli to be learned are random, uncorrelated and each of them is presented once. They pointed out the importance of (i) the stochasticity of the learning dynamics, (ii) the sparseness of the stimuli, and (iii) the global balance between numbers of potentiating and depressing transitions. Amit and Brunel (1995) used simulations to study such a learning process in a more realistic case in which stimuli are drawn randomly at each presentation from a set of classes, defined by clusters in stimulus space. The differences in the patterns of activity evoked by stimuli belonging to the same class may reflect either some noise added to the signal in the visual pathways (early preprocessing) or small differences in some of the features of the visual stimuli. For example, the stimulus degradation studied by Amit *et al* (1997) leads to gradual modifications in the visual response that are observable at the level of the inferotemporal cortex of a behaving monkey. Amit and Brunel (1995) demonstrated how the resulting modifications in the synaptic structure could stabilize attractors (internal representations) corresponding to the extracted prototypes of the classes of shown stimuli. The existence of an attractor state (a stable pattern of activities) correlated with a prototype defines whether the corresponding class is learned or not, since it makes it possible to maintain an active representation of this class, in the absence of the stimulus correlated with it. This is the definition of learning adopted in this paper.

The present study represents an analytical approach to this learning process. The environment, from which the network learns, is defined by a set of prototypes. Each prototype defines a class, or cluster, of similar patterns of activities (patterns belonging

to a class are correlated with the corresponding prototype). At discrete points in time, a randomly selected pattern is presented and synaptic transitions take place. The sequence of presentations of stimuli is composed of members of these classes, in which at each presentation both the prototype and the specific class member are chosen at random. It is a relatively common protocol of visual memory experiments. The environment (i.e. the set of prototypes or classes) is either fixed, or changing. In the latter case, classes are added to or removed from the set. This allows the study of the speed at which classes are learned or forgotten.

Any presentation of a stimulus provokes stochastic transitions between synaptic efficacies. Each synapse has two stable states, a low background state and an elevated state. Transitions from low to high are functionally similar to LTP, while the reverse transitions are similar to LTD. The model can incorporate different types of LTD, such as homosynaptic, heterosynaptic, or both. Such a model of unsupervised synaptic plasticity can be easily implemented in a material device (Badoni *et al* 1995, Annunziato 1995).

In Amit and Fusi (1994) the memory capacity of this model had been estimated by assuming that each pattern can be learned after a single presentation. By contrast, in some areas in which attractors are observed, e.g. inferotemporal (IT) cortex, a large number of presentations of the same pattern seem to be required to produce reverberating activity (Miyashita 1993), which suggests a rather slow learning process. In this paper we shall study such a *slow learning* situation, i.e. a scenario in which a given pattern has to be shown many times before it is learned.

1.1. Summary of the main results

The main results in the slow learning limit are as follows.

1.1.1. Dependence on the sequence of presentations. In the slow learning limit and if there are no temporal correlations in the sequence (i.e. choices of the class to be presented at any two different times are uncorrelated), the degree of correlation between the synaptic matrix and a given prototype, which determines whether this prototype is learned or not, does not depend on the specific sequence of presentations of stimuli but only on the set of stimuli in the environment. The final configuration of the synaptic matrix contains information about the statistics of the last stimuli presented, those appearing in a sliding time window whose length depends on synaptic transition probabilities. For a fixed environment, as the transition probability goes to zero, this time window gets longer and (i) the number of times a given prototype appears in it increases; (ii) for a given presentation, the number of modified synapses decreases. This means that, following any presentation, a smaller amount of information is acquired, and also that a smaller fraction of synapses forgets the patterns presented in the past. The final outcome is a synaptic matrix which is less biased toward the most recent stimuli compared to conventional palimpsest memories (Nadal *et al* 1986, Parisi 1986). Nevertheless, the system is still able to adapt itself to new environments since the sliding time window is finite. The price to be paid is a longer adaptation time: as the transition probability decreases the adaptation time increases.

1.1.2. Categorization properties. If learning is slow enough, the memory of the class prototypes is always stronger than the memory of any other members of the same class shown. Thus, the learning process naturally categorizes the input stimuli by extracting a good representative of a class of similar patterns in the stream of stimuli. If the members are generated as clouds around fixed prototypes, the extracted representatives become closer

and closer to the real prototypes as the sliding memory window becomes larger, i.e. when the transition probabilities get smaller. Intuitively, a larger memory window means a larger sampling of the entire class. In this paper we focus on the limit of very small transition probabilities, and we show that in this limit the extracted prototypes coincide with the prototypes underlying the stream of stimuli. If the environment changes and the prototypes move in the space of all possible activity configurations, then the synaptic configuration is able to adapt to new prototypes, provided that they move slowly enough.

1.1.3. Storage capacity. This is defined as the maximum number of classes that can be memorized. It depends on the sparseness of the internal representations of the stimuli (the mean fraction of neurons activated by each stimulus): the sparser the representation, the larger the storage capacity. A good storage capacity also requires a global balance between LTP and LTD, as in (Amit and Fusi 1994). Moreover, we show that it is independent of the specific implementation of LTD provided the global balance between LTP and LTD is preserved.

1.1.4. Learning and forgetting rates. We define the learning rate as the inverse of the typical number of repetitions of a new class which must be presented in order for it to be learned. The time at which a new class becomes recallable depends on this number of repetitions and on the frequency with which the network is presented the stimuli of the environment.

We also consider a situation in which a class is removed from the environment and the network is still presented a stream of uncorrelated stimuli belonging to other classes. In this case, the removed class is forgotten after a certain number of presentations of each remaining class. The forgetting rate is defined as the inverse of this number of presentations.

Learning and forgetting rates are proportional to the LTP/LTD transition probabilities. The learning rate is almost independent of the number of classes present in the environment, except when it is close to the storage capacity. Near the limit of storage capacity it decreases sharply, owing to the fact that all the information about a specific class is forgotten in the interval between two successive presentations of the members of the class. The forgetting rate is much smaller than the learning rate when there are few classes in the environment, but increases as the number of classes increases, until it eventually becomes larger than the learning rate. The forgetting rate becomes infinite when the storage capacity is reached.

These results are demonstrated both analytically and by simulations. The organization of the paper is as follows: first we define the model neurons and the external stimuli. In section 2 we define the synaptic dynamics. In section 3 we calculate the synaptic distribution for a generic sequence and in section 4 we study the slow learning situation. Then in section 5 we study the sparse coding limit. This allows us to determine relatively simple expressions for the quantities of interest. Lastly, in section 6 we describe the results of the simulations that confirm the main analytical results.

2. The model

2.1. Neuronal response to stimuli

We consider a network composed of a large number of neurons, which are taken to represent the pyramidal cells of a cortical network. Each neuron in the network is labelled by an index $i = 1, \ldots, N$ where N is the number of neurons. In this paper we do not consider the

neuronal dynamics explicitly, but rather consider the steady state imposed on the network by an external stimulus. For simplicity, any stimulus leaves neuron i at one of two possible activity states, V_i $i = 0, 1$ which may be related to spike rates: $V_i = V_0$ the neuron shows no visual response (spontaneous activity state); $V_i = V_1 \gg V_0$ the neuron has visual response. The ensemble of activations in the network by a given stimulus represents the way in which this stimulus is encoded in the network.

A stimulus shown at time t can therefore be characterized by a binary string $\{\xi_i^t = 0, 1\}$, which determines the activation of all neurons in the network during its presentation:

$$V_i(t) = \begin{cases} V_1 & \text{if } \xi_i^t = 1 \\ V_0 & \text{if } \xi_i^t = 0. \end{cases}$$

The population of neurons which is activated by a given stimulus will be called the foreground of this stimulus. The remaining neurons define its background.

2.2. Statistics of stimuli: classes

Stimuli shown to the network belong to a set of p predetermined classes, which defines the 'environment' of the network. Each class is defined by a representative pattern, the *prototype* η^μ, $\mu = 1, \ldots, p$. Prototypes are random and uncorrelated N-bit words chosen according to

$$\Pr(\eta_i^\mu = 1) = f \qquad \Pr(\eta_i^\mu = 0) = 1 - f \tag{1}$$

where f is the coding level (or sparseness) of the class prototypes.

Each prototype defines its corresponding class (or cluster) of stimuli. The members of a class are noisy versions of the prototype. In a visual memory experiment the noise may be interpreted as follows.

- Noise due to preprocessing in the early visual stages or to small eye movements: even if the animal is always presented the same stimulus, the pattern of activity in the network might be different from presentation to presentation because of the noise generated by other networks.
- Degradation of the visual stimulus: the patterns of activities corresponding to the members of a specific class, are the outcome of the degradation of the visual stimulus (e.g. when RGB noise is added to the prototype, as in Amit *et al* (1997)).
- Small changes in the visual stimuli (e.g. in one or more of the features) that induce modifications in the visual response of the neurons of the network. The similarity of the visual stimuli is reflected by the correlations between the prototype and stimuli belonging to the same class.

A stimulus $\eta^{\mu\nu}$ belonging to class μ is chosen randomly in the following way.

- If the neuron is in the foreground of the prototype, $\eta_i^\mu = 1$:

$$\Pr(\eta_i^{\mu\nu} = 1) = 1 - x(1 - f) \qquad \Pr(\eta^{\mu\nu} = 0) = (1 - f)x. \tag{2}$$

- If the neuron is in the background of the prototype, $\eta_i^\mu = 0$:

$$\Pr(\eta^{\mu\nu} = 1) = fx \qquad \Pr(\eta^{\mu\nu} = 0) = 1 - fx. \tag{3}$$

x measures the extent of a given class or the distance between a typical instance and its class prototype. If $x = 0$, instances are identical to their prototype. If $x = 1$, examples are uncorrelated with their prototype. This procedure ensures that the average fraction of activated neurons is f.

A sequence of presentations specifies which stimulus is shown at each time step t. In a random sequence, a class is selected at random at each presentation with equal probability $1/p$, and then an instance is generated at random from the class using equations (2) and (3). The formalism developed in this paper could easily be generalized to the case in which the probabilities of presentation are different from class to class. Choices of the class and of the class member at two different presentations are uncorrelated. The case of temporal correlations in the sequence of presentations will be considered in a future study (see also (Brunel 1996) for a study of learning of temporal correlations in a simplified situation in which stimuli activate non-overlapping sets of neurons).

In the following, ξ^t always denotes a generic stimulus shown at time t, η^μ a class prototype, $\eta^{\mu\nu}$ an instance ν of class μ.

2.3. Synaptic dynamics and the standard learning model

Two essential features characterize the synaptic dynamics: the fact that each synapse is discrete with a finite number of states (limited analogue depth); and the stochasticity of transitions occurring from state to state. In a material neural network the elements are assumed to be implementable in simple devices (electronic or biochemical). It is unlikely that a biological or an electronic device could preserve a large set of stable values during long time intervals, in the absence of a stimulus.

In what follows we assume that on long timescales the synaptic efficacy has only two stable states because:

(i) such a structure is sufficient for learning selective delay activity;
(ii) it simplifies the calculations and the design of the electronic implementation of the synapse;
(iii) the discreteness of synaptic efficacies is consistent with experiment (Bliss and Collingridge 1993).

The synaptic efficacy connecting neuron j to neuron i is denoted by J_{ij}. All synaptic efficacies have two stable states, denoted by $J_0 < J_1$. In the following, for simplicity, we use $J_1 = 1$, $J_0 = 0$, without loss of generality.

Stochastic transitions between these states may occur if either the presynaptic or the postsynaptic neuron is active during presentation of a stimulus. At time t, when a stimulus $\{\xi_i^t\}$ is presented:

- If $J_{ij} = 0$, then $J_{ij} \to 1$ with probability $a(\xi_i^t, \xi_j^t)$ (LTP transition). $a(\xi_i^t, \xi_j^t)$ can be written as

$$a(\xi_i^t, \xi_j^t) = q_+ p(\xi_i^t, \xi_j^t)$$

  where q_+ is the intrinsic potentiation probability, and $p(\xi_i^t, \xi_j^t) \in [0, 1]$ is a function carrying the dependence on the activities of the two neurons connecting the synapse. If the two activities ξ_i^t, ξ_j^t are such that $p = 1$, then an LTP transition occurs with probability q_+. Otherwise $p = 0$ and no LTP transition can occur.
- If $J_{ij} = 1$, then $J_{ij} \to 0$ with probability $b(\xi_i^t, \xi_j^t)$ (LTD transition). Again, we can write

$$b(\xi_i^t, \xi_j^t) = q_- d(\xi_i^t, \xi_j^t)$$

  where q_- is the intrinsic depression probability, and $d(\xi_i^t, \xi_j^t) \in [0, 1]$.

Thus, learning is a random walk among the two stable states of synaptic efficacy, and any synapse is characterized at any time t by a probability distribution ($\Pr(J_{ij} = 1, t)$,

$\Pr(J_{ij} = 0, t)$), which depends on all stimuli which have been presented to the network before time t.

The formalism developed in this paper applies to any functions p and d describing LTP and LTD. In most of the following, however, we shall restrict ourselves to a symmetric learning dynamics introduced in (Amit and Fusi 1994, Amit and Brunel 1995) which captures features of experimental results on synaptic plasticity (Bliss and Collingridge 1993, Christie *et al* 1994):

$$p(\xi_i^t, \xi_j^t) = \xi_i^t \xi_j^t \tag{4}$$

corresponding to LTP obtained only when both presynaptic and postsynaptic neurons are sufficiently depolarized; and

$$d(\xi_i^t, \xi_j^t) = \xi_i^t(1 - \xi_j^t) + \xi_j^t(1 - \xi_i^t) \tag{5}$$

corresponding to LTD in the case of either postsynaptic but not presynaptic activity (heterosynaptic LTD), or presynaptic but not postsynaptic activity (homosynaptic LTD). This learning dynamics will be referred to as the standard learning model (SLM) in the following. In section 5.3 we shall compare different types of LTD characterized by different mixtures of homosynaptic and heterosynaptic LTD:

$$d(\xi_i^t, \xi_j^t) = u\xi_i^t(1 - \xi_j^t) + v\xi_j^t(1 - \xi_i^t) \tag{6}$$

where u (v), varying in the interval $[0, 1]$, is the relative strength of heterosynaptic (homosynaptic) LTD.

2.4. Measures of the degree of learning

To monitor the effect of learning on the synaptic matrix we define the following quantities.

(i) The mean potentiation g at time t:

$$g(t) = \frac{1}{N(N-1)} \sum_{i \neq j} J_{ij}(t) \tag{7}$$

is the average fraction of potentiated synapses in the network at time t.

(ii) The mean 'intra-class' (ICP) potentiation g^μ of the prototype of class μ (Amit and Brunel 1995)

$$g^\mu(t) = \frac{1}{fN(fN-1)} \sum_{i,j} J_{ij}(t)\eta_i^\mu \eta_j^\mu \tag{8}$$

measures the correlation between the prototype of class μ and the synaptic matrix, and thus of the degree of learning of the corresponding class. If $g^\mu = g$ there is no correlation between the class and the synaptic matrix.

(iii) The mean 'intra-example' (IEP) potentiation $G^{\mu\nu}$ of instance ν of class μ:

$$g^{\mu\nu}(t) = \frac{1}{fN(fN-1)} \sum_{i,j} J_{ij}(t)\eta_i^{\mu\nu} \eta_j^{\mu\nu} \tag{9}$$

measures the correlation between a particular instance and the synaptic matrix. It can be used to evaluate whether the synaptic matrix is more correlated with instances or with their class prototypes.

Realistic associative memory models (see e.g. Amit and Brunel 1995, 1997) show that, depending on the value of g_+ relative to g, two types of behaviour are observed. When $g_+ - g < \Delta g_c$, where Δg_c depends on the details of the neuronal model, the network is unable to sustain attractors correlated with learned stimuli. In this case, the only stable state in the network is a completely silent state or a spontaneous activity state in a network of spiking neurons with strong recurrent inhibition. On the other hand, when the ICP becomes significantly larger than the mean potentiation in the network, network states correlated with the class prototype become stable. The precise criteria depend on the neuronal model; for the particular case of a network of analogue neurons characterized by a continuous transfer function we have typically $\Delta g_c \sim 0.5$.

On the other hand, the relative values of the IEP and the ICP determine whether the attractor stabilized by the network is more correlated with the class prototype (thus categorizing the input stimuli) or with the most recent instance of the class.

3. Synaptic distribution for a generic sequence

3.1. Sources of stochasticity: learning as a Markov process

The intrinsic stochastic mechanism which drives transitions from one stable state to another is not the only source of stochasticity. Learning can be considered stochastic either due to the dynamics of the synaptic modifications or due to the nature of the data shown to the network (Amit and Fusi 1994, Heskes and Kappen 1991). In fact, the sequence of random uncorrelated stimuli presented to the network represents a sequence of random activity levels imposed on the synapse. Since the learning dynamics is local in time, the presentation of a sequence of uncorrelated random stimuli induces a Markov process on the set of values of each synapse. In our case the synapse has only two stable states, so the synaptic dynamics can be fully described in terms of the transition probabilities of transferring to the up/down state.

3.2. Probability distribution as an explicit function of the sequence of stimuli

We start our analysis with the computation of the final synaptic distribution as an explicit function of the activities η_i^μ presented to the synapse by each stimulus of the sequence. The transition matrix corresponding to the Markov process of a single presentation at time t can be written as

$$M_{ij}(t) = \begin{pmatrix} 1 - a_{ij}^t & a_{ij}^t \\ b_{ij}^t & 1 - b_{ij}^t \end{pmatrix}$$

where $a_{ij}^t = a(\xi_i^t, \xi_j^t)$ ($b_{ij}^t = b(\xi_i^t, \xi_j^t)$) is the probability that synapse J_{ij} is potentiated (depressed) following the presentation of pattern ξ^t at time t, as defined in section 2.3.

From this transition matrix, we can find the probability distribution of the synaptic efficacies at time T, as a function of the initial distribution and of all stimuli presented between 0 and T.

The probabilities of the synapse ij being in the excited state ($G_{ij}(T) \equiv \Pr(J_{ij} = 1, T)$) or in the background state ($1 - G_{ij}(T) \equiv \Pr(J_{ij} = 0, T)$) at time T are calculated explicitly in appendix A. We obtain:

$$G_{ij}(T) = \sum_{t=1}^{T} a_{ij}^t \prod_{s=t+1}^{T} \lambda_{ij}^s + G_{ij}(0) \prod_{s=1}^{T} \lambda_{ij}^s \tag{10}$$

in which $\lambda_{ij}^t = 1 - a_{ij}^t - b_{ij}^t$. λ_{ij}^t appears in equation (10) as an 'instantaneous decay factor' at time t: it is the probability that no transition occurred at time t, given that stimulus ξ^t had been presented. In equation (10), each term in the sum on the right-hand side (RHS) corresponds to a stimulus presented at time $t < T$. These terms are weighted by the 'decay' factor $\prod_s \lambda^s$ provoked by successive presentations. This factor is the probability that no synaptic transition occurred between time t and time T. In this way, the contributions of earlier stimuli are obscured by more recent ones. If there is a finite probability that potentiating and (or) depressing stimuli occur in the sequence, after a sufficiently long time ($T \to \infty$) the second term in the RHS of equation (10) vanishes, i.e. the synapse forgets its initial condition, and we obtain

$$G_{ij}(T) = \sum_{t \leqslant T} a_{ij}^t \prod_{s=t+1}^{T} \lambda_{ij}^s. \tag{11}$$

This is the case in which at least one of the prototypes in the 'environment' tends to potentiate or depress the synapse, or where the size of clusters around the prototypes is finite ($x > 0$).

Note that the synaptic distribution, equation (11), depends, in the case of the presentation protocol described in section 2.2, on

- the set of classes;
- the particular realization of the (random) sequence of presentation of classes, i.e. the specification of the classes shown at each time step;
- and, in the case of classes of finite extent, the extraction of class members at each time step.

Thus the distribution of synaptic efficacies depends on the specification of the entire sequence of stimuli. On the other hand, one would like to know the properties of the synaptic matrix for a 'typical' sequence. We shall see in section 4 that this is possible when the transition probabilities q_+, q_- are small: in this case the *average* over all possible realizations of the sequence of presentations gives a good approximation for most realizations, in the sense that the variability of the average potentiation levels from sequence to sequence goes to zero with the transition probabilities.

4. Slow learning: averaging over random sequences

To study the properties of the synaptic matrix in the slow learning scenario we shall proceed as follows.

(i) We start from the general expression for the synaptic distribution following the presentation of a specific, arbitrary sequence of stimuli (equation (10)). This is an explicit function of the pairs of activities (ξ_i^t, ξ_j^t) imposed on the synapse by all the stimuli ξ_i^t $(0 < t < T)$ of the input stream.
(ii) The synaptic distribution is then averaged over all possible realizations of the random sequences defined in section 2.2.
(iii) Next we calculate the average over all sequences of the potentiation levels g and g^μ of equations (7) and (8).
(iv) Finally we compute the variability of these potentiation levels from sequence to sequence in section 4.3, and find that when learning is sufficiently slow, this variability becomes negligible with respect to the average potentiation levels.

Hence, in the slow learning limit the sequence-averaged potentiation levels give a good estimate of the statistical properties of a 'typical' synaptic matrix which has seen a random sequence of the set of stimuli.

4.1. Sequence average of synaptic efficacy

The average of a quantity over all possible sequences is denoted by $\langle \ldots \rangle$. For a random sequence, when only prototypes are shown, ξ_i^t is chosen at random in the set of all the prototypes independently at all t: $\xi_i^t = \eta_i^\mu$ with probability $1/p$ for each $\mu = 1, \ldots, p$. We have, for example,

$$\langle \xi_i^t \rangle = \frac{1}{p} \sum_{\mu=1}^{p} \eta_i^\mu .$$

In order to perform the average over sequences of the synaptic distribution, (11), we note that each of the $a_{ij}^t \prod_s \lambda_{ij}^s$ is a product of terms corresponding to different times, and that they can be averaged independently since presentations at different time steps are uncorrelated. Thus we obtain:

$$\langle G_{ij} \rangle = \langle a_{ij} \rangle \sum_{s=0}^{\infty} \langle \lambda_{ij} \rangle^s = \frac{\langle a_{ij} \rangle}{\langle a_{ij} + b_{ij} \rangle} \tag{12}$$

in which

$$\langle a_{ij} \rangle = \frac{q_+}{p} \sum_\mu p(\eta_i^\mu, \eta_j^\mu) \qquad \langle b_{ij} \rangle = \frac{q_-}{p} \sum_\mu d(\eta_i^\mu, \eta_j^\mu). \tag{13}$$

Defining

$$P_{ij} = \sum_\mu p(\eta_i^\mu, \eta_j^\mu) \qquad D_{ij} = \sum_\mu d(\eta_i^\mu, \eta_j^\mu) \tag{14}$$

we obtain the sequence-averaged probability that synapse J_{ij} is potentiated,

$$g_{ij} = \langle G_{ij} \rangle = \frac{q_+ P_{ij}}{q_+ P_{ij} + q_- D_{ij}}. \tag{15}$$

4.2. Sequence-averaged potentiation levels

To calculate the sequence-averaged potentiation levels, equations (7)–(9), we simply have to replace J_{ij} in these equations by its sequence-average g_{ij} given by equation (15). In this way, we obtain the sequence-averaged potentiation level in the network,

$$g = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{q_+ P_{ij}}{q_+ P_{ij} + q_- D_{ij}} \tag{16}$$

and the sequence-averaged ICP for a generic class μ:

$$g^\mu = \frac{1}{fN(fN-1)} \sum_{i \neq j} \frac{q_+ P_{ij}}{q_+ P_{ij} + q_- D_{ij}} \eta_i^\mu \eta_j^\mu . \tag{17}$$

The sum on the RHS of equation (16) can be replaced by a sum over all possible values of P_{ij} and D_{ij} weighted by their joint probability distribution $\psi(\Pi, \Delta) = \Pr(P_{ij} = \Pi, D_{ij} = \Delta)$; in these terms g becomes

$$g = \sum_{\Pi, \Delta = 0}^{p} \psi(\Pi, \Delta) \frac{q_+ \Pi}{q_+ \Pi + q_- \Delta}. \tag{18}$$

Equation (17) can be rewritten in a similar way as a sum over all possible values of P_{ij} and D_{ij} weighted by the joint probability distribution conditional on the synapse experiencing coincident activation of presynaptic and postsynaptic neurons when a generic prototype μ is presented, $\psi_+(\Pi, \Delta) = \Pr(P_{ij} = 1 + \Pi, D_{ij} = \Delta | \eta_i^\mu \eta_j^\mu = 1)$. The average intra-class potentiation level is thus

$$g_+ = \sum_{\Pi=0}^{p-1} \sum_{\Delta=0}^{p} \psi_+(\Pi, \Delta) \frac{q_+(1+\Pi)}{q_+(1+\Pi) + q_-\Delta}. \tag{19}$$

Notice that since all classes have the same probability of being shown at any time step, g_+ does not depend on μ. The joint distributions $\psi(\Pi, \Delta)$ and $\psi_+(\Pi, \Delta)$ depend on the parameters defining the stimuli, i.e. f and p, and on the particular learning dynamics, defined by the Boolean functions p and d.

4.3. Variability from sequence to sequence

The asymptotic variability in the potentiation level is defined by

$$\Delta g^2 = \frac{1}{N^2(N-1)^2} \sum_{i \neq j, k \neq l} \langle \Delta G_{ij} \Delta G_{kl} \rangle \tag{20}$$

where $\Delta G_{ij} = G_{ij} - \langle G_{ij} \rangle$. The variability in the ICP is defined in a similar way:

$$\Delta g_\mu^2 = \frac{1}{f^2 N^2 (fN-1)^2} \sum_{i \neq j, k \neq l} \langle \Delta G_{ij} \Delta G_{kl} \rangle \eta_i^\mu \eta_j^\mu \eta_k^\mu \eta_l^\mu. \tag{21}$$

When the transition probabilities $q_+ \sim q, q_- \sim \chi q$ and q go to zero, both variabilities go to zero with q as (see appendix B for details):

$$\Delta g^2 = q \frac{1}{N^2(N-1)^2} \sum_{i \neq j, k \neq l} I_{ijkl} \qquad \Delta g_\mu^2 = q \frac{1}{f^2 N^2 (fN-1)^2} \sum_{i \neq j, k \neq l} I_{ijkl} \eta_i^\mu \eta_j^\mu \eta_k^\mu \eta_l^\mu \tag{22}$$

where I_{ijkl}, given in appendix B, goes to a finite quantity when $q \to 0$. If q is small enough, the variability from sequence to sequence becomes negligible with respect to the corresponding averages, which remain finite in this limit, and the calculation of the sequence-averaged quantities gives a good estimate of the corresponding quantity after a typical sequence has been shown. In the following, we shall focus on sequence-average quantities only.

4.4. Speed of learning and forgetting

If at time $t = 0$ the synaptic distribution is $G_{ij}(T = 0) = G_0$ for all (i, j), the distribution at time $T > 0$ is given by equation (10), i.e.

$$G_{ij}(T) = G_0 \prod_{s=1}^{T} \lambda_{ij}^s + \sum_{t=1}^{T} a_{ij}^t \prod_{s=t+1}^{T} \lambda_{ij}^s. \tag{23}$$

The first term on the RHS describes the decay of the initial condition G_0, or the forgetting of whatever was learned before $t = 0$, while the second term describes learning after $t = 0$. If one waits for a sufficiently long time, all memory of the past will be erased and one arrives at the asymptotic expression, equation (10).

Taking once more the average over sequences we obtain

$$g_{ij}(T) = \frac{\langle a_{ij} \rangle}{\langle a_{ij} + b_{ij} \rangle} + \langle \lambda_{ij} \rangle^T \left(G_0 - \frac{\langle a_{ij} \rangle}{\langle a_{ij} + b_{ij} \rangle} \right). \tag{24}$$

The first term in equation (24) corresponds to the asymptotic distribution, equation (12). The second term is the product of the 'decay' term $\langle\lambda_{ij}\rangle^T$ by the difference between the initial distribution and the asymptotic one. This second term describes the response of the system when there is a change in the statistics of the input. Each synapse has an associated decay time constant $\tau_{ij} = -1/\log\langle\lambda_{ij}\rangle$. To describe the population-averaged behaviour in response to such changes it is useful to define a 'forgetting' and a 'learning' function.

The forgetting function describes the evolution of the intra-class potentiation of a prototype which is perfectly learned (i.e. the intra-class connectivity has reached its asymptotic value g_+) at time $T = 0$ and is not shown any more for $T > 0$, assuming that it is uncorrelated with stimuli that are presented for $T > 0$,

$$\phi(T) = g + \frac{1}{N(N-1)}\sum_{i\neq j}\langle\lambda_{ij}\rangle^T\left(g_+ - \frac{\langle a_{ij}\rangle}{\langle a_{ij}+b_{ij}\rangle}\right). \tag{25}$$

This function is such that $\phi(t=0) = g_+$ and $\phi(t\to\infty) = g$.

The 'learning' function of prototype μ describes how a prototype, uncorrelated with the initial synaptic distribution $G_{ij}(0)$, is learned as a function of time (i.e. the evolution of its ICP), assuming that the initial distribution of intra-class synapses is equal to the asymptotic connectivity, $G_0 = g$

$$\phi^\mu(T) = g_+ + \frac{1}{fN(fN-1)}\sum_{i\neq j}\langle\lambda_{ij}\rangle^T\left(g - \frac{\langle a_{ij}\rangle}{\langle a_{ij}+b_{ij}\rangle}\right)\eta_i^\mu\eta_j^\mu. \tag{26}$$

It is defined such that $\phi_+(T=0) = g$ and $\phi_+(T\to\infty) = g_+$.

We can again express these functions in terms of the joint probabilities of P_{ij} and D_{ij}. In these terms, the forgetting function is

$$\phi(T) = g + \sum_{\Pi,\Delta=0}^{p}\psi(\Pi,\Delta)\left(g_+ - \frac{q_+\Pi}{q_+\Pi+q_-\Delta}\right)\left(1-\frac{1}{p}(q_+\Pi+q_-\Delta)\right)^T \tag{27}$$

and the learning function is

$$\phi_+(T) = g_+ + \sum_{\Pi=0}^{p-1}\sum_{\Delta=0}^{p}\psi_+(\Pi,\Delta)\left(g - \frac{q_+(1+\Pi)}{q_+(1+\Pi)+q_-\Delta}\right)\left(1-\frac{1}{p}(q_+(1+\Pi)+q_-\Delta)\right)^T. \tag{28}$$

4.5. An example: SLM, random sequence of prototypes only

In the SLM described in section 2.3, if we present, in a random way the class prototypes only, the joint probability distribution $\psi(\Pi,\Delta)$ is

$$\psi(\Pi,\Delta) = \frac{p!}{\Pi!\Delta!(p-\Pi-\Delta)!}2^\Delta f^{2\Pi+\Delta}(1-f)^{2(p-\Pi)-\Delta}.$$

Similarly, ψ_+ is given by

$$\psi_+(\Pi,\Delta) = \frac{(p-1)!}{\Pi!\Delta!(p-1-\Pi-\Delta)!}2^\Delta f^{2\Pi+\Delta}(1-f)^{2(p-1-\Pi)-\Delta}$$

for $0 \leqslant \Pi+\Delta \leqslant p-1$.

Equations (18), (19), (27) and (28) then determine the statistical properties of the synaptic matrix, in both stationary and non-stationary environments, as a function of the transition probabilities q_+ and q_-, and of the parameters characterizing the flow of stimuli, f and p.

For any $f > 0$ and finite p, $g_+ > g$. In the limit $p \to \infty$, $g_+ \to g$, i.e. the synaptic matrix becomes uncorrelated with the shown prototypes when their number grows to infinity. Imposing a finite degree of correlation between the synaptic matrix and class prototypes, $g_+ - g > \Delta g_c$, implies a finite upper bound on the maximal number of classes that can be learned.

On the other hand, in the next section it is shown that the maximal number of learnable prototypes grows to infinity when the sparseness f goes to zero.

5. Study of the SLM in the sparse coding limit

Visual memory experiments in performing monkeys indicate that neuronal patterns of activity sustained in IT cortex during the delay period of such experiments involve a low fraction of neurons (see e.g. Miyashita 1988), and lead to the conclusion that coding levels in this area are of the order of 1–2% (Brunel 1994). This suggests that one should focus on the limit of a low coding level (sparse coding limit). Previous studies, both with fixed synapses (see e.g. Meunier and Nadal 1995 and references therein) and with dynamic synapses in the case of one-shot learning of uncorrelated stimuli (Amit and Fusi 1994), have shown that the information capacity of the system increases sharply when f decreases.

In the limit $f \to 0$, to leading order in f, the variables Π and Δ become independent and

$$\psi(\Pi, \Delta) = \psi_p(\Pi)\psi_d(\Delta) + O(f)$$

(see appendix C), in which ψ_p (ψ_d) is the distribution of the number of prototypes that tend to potentiate (depress) the synapse. Furthermore, if p goes to infinity as f goes to zero, the distributions ψ_+ and ψ become more and more similar, and we have

$$\psi_+(\Pi, \Delta) = \psi_p(\Pi)\psi_d(\Delta) + O(f).$$

In this limit, ψ_p and ψ_d become Poisson distributions,

$$\psi_p(\Pi) = (pf^2)^\Pi \frac{\exp(-pf^2)}{\Pi!} \qquad \psi_d(\Delta) = (2pf)^\Delta \frac{\exp(-2pf)}{\Delta!}. \tag{29}$$

The behaviour of ψ_p and ψ_d depends on how the number of classes p scales with f.

- If $p \ll 1/f$ the probability of seeing (at least) one potentiating (depressing) prototype is of the order of pf^2 (pf). Thus, most synapses never see either a potentiating or depressing prototype.
- If $p \sim 1/f$ the probability of seeing a potentiating prototype is of the order of f, but now any synapse will typically see a few depressing prototypes, as given by the Poisson distribution, equation (29).
- If $1/f \ll p \ll 1/f^2$ the probability of seeing a potentiating prototype is still very small, but on the other hand the distribution of the number of depressing prototypes becomes Gaussian with mean pf and variance pf.
- If $p \sim 1/f^2$ a synapse typically experiences a finite number of potentiating prototypes, as given by the Poisson distribution, equation (29).
- If $p \gg 1/f^2$ both distributions become Gaussian. Synapses see a large number of potentiating and depressing prototypes.

Thus, we have two crossover regimes, $p \sim 1/f$ (hereafter called low-loading regime) and $p \sim 1/f^2$ (high-loading regime). For $f \sim 0.01$, the low-loading regime corresponds to a number of classes p of the order of 100, while the high-loading regime corresponds to p of the order of 10 000.

In the following q_+ is low but stays finite when $f \to 0$, q_- is of the order of f, so that the number of potentiations and depressions is of the same order. We denote

$$q_+ = q \qquad q_- = \rho f q. \tag{30}$$

This scaling of q_-/q_+ with f is optimal in terms of the number of classes the system is able to memorize. If q_- stays finite as f goes to zero, the system is able to store and recall up to $p \sim 1/f$ classes, because if $p \gg 1/f$ most synapses, including the 'intra-class' synapses, will see a very large number of depressing prototypes and will be depressed. If $q_- \sim f$ the system can store up to $p \sim 1/f^2$ classes. If $q_- \ll f$ the system becomes useless as a memory device since most synapses will become potentiated, as in the Willshaw model above its critical capacity (Willshaw *et al* 1969). In the following we shall only mention the results in the text. The details of the calculations can be found in appendix D.

5.1. Learning pure prototypes

5.1.1. Low loading. We take $p = \alpha/f$, where α is a finite parameter, and $f \to 0$. The average potentiation level is simply

$$g = g_0 \exp(-2\alpha)$$

where g_0 is the initial potentiation level. In this case, the synaptic matrix keeps a memory of the initial condition. This memory decreases as α increases. This means that showing a low number, $p \ll 1/f$, of prototypes for an arbitrary long time will not erase the memory of whatever was learned before, simply because most synapses will not make transitions. This is true only for pure prototypes. As soon as class members differ from the prototype, no memory of the initial conditions will persist, as we show in the following. When α becomes large (but $\alpha f \ll 1$), g goes to zero.

The average IC potentiation level is $g_+ = 1 - \mathrm{O}(f)$. In this regime all classes are perfectly learned in the sparse coding limit.

The forgetting function

$$\phi(T) = \exp\left[-2\alpha\left(1 - \exp\left[-\frac{qf^2\rho T}{\alpha}\right]\right)\right]$$

for T small compared with $\alpha/(q\rho f^2)$, satisfies

$$\phi(T) \sim \exp(-2qf^2\rho T).$$

Thus, forgetting occurs with a time constant $\sim 1/(2q\rho f^2)$. This is simply the time constant of LTD in the network, i.e. the inverse of the probability that a synapse is depressed when a random pattern is shown.

The learning function is

$$\phi_+(T) = 1 + (g-1)\exp\left(-\frac{qfT}{\alpha}\right).$$

Learning occurs with a time constant $\sim \alpha/(qf)$. This is simply the product of the number of prototypes p multiplied by the LTP rate (i.e. $1/q$) for an intra-class synapse. There is a factor $1/f$ between learning and forgetting timescales and thus when $f \to 0$ learning is much faster than forgetting.

High loading. We consider now the case $p = \alpha/f^2$. The average potentiation level is now given by equation (D6). It increases from $g = 0$ when $\alpha = 0$ to

$$g = \frac{1}{1+2\rho}$$

when α goes to infinity. This is the value it would have if a flow of uncorrelated stimuli had been presented to it.

The average IC potentiation level, equation (D7) is simply related to g by the expression

$$g_+ = 1 - 2\rho g.$$

It decreases from 1 at $\alpha = 0$ to $1/(1+2\rho)$ when α goes to infinity. Thus, in this limit $g_+ \to g$ and the synaptic matrix becomes uncorrelated with the shown prototypes. If the neuronal dynamics is such that the criterion for memory retrieval is $g_+ - g > \Delta g_c = 0.5$, as in the model studied in (Amit and Brunel 1995), we obtain that the maximal number of classes that can be learned, for $\rho = 1$, is $p \sim 0.3/f^2$. Taking again $f \sim 0.01$ we find that up to about 3000 classes can be learned in the synaptic matrix.

The forgetting and learning functions are given by equations (D8) and (D9). From these expressions we can calculate the number of presentations needed to learn a new prototype or to forget an old one. The time τ_+ to learn a new prototype is given by imposing

$$\phi_+(\tau_+) = g + \Delta g_c.$$

The time τ to forget an old prototype is given by:

$$\phi(\tau) = g + \Delta g_c.$$

To obtain a quantitative idea about the learning and forgetting timescales suppose that $f = 0.01$, $\rho = 1$, and $q = 0.002$. If the environment consists of 1000 classes, it follows that to learn a newly presented class we must wait about $\tau_+ = 400\,000$ presentations (400 per class), while to forget a class which is not presented any more we must wait about 10^6 presentations. On the other hand, if the environment is composed of 100 classes, we need

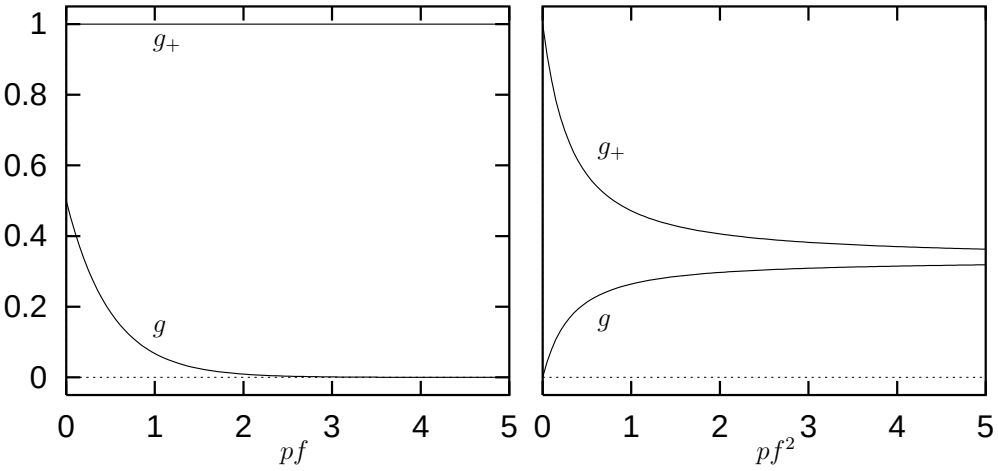

Figure 1. Average levels of potentiation (g) and of intraclass potentiation (g_+) as a function of the number of prototypes p in the sparse coding limit. Left: low loading ($p \sim 1/f$), $g_0 = 0.5$. Right: high loading ($p \sim 1/f^2$). The quality of learning of the shown classes degrades as pf^2 increases.

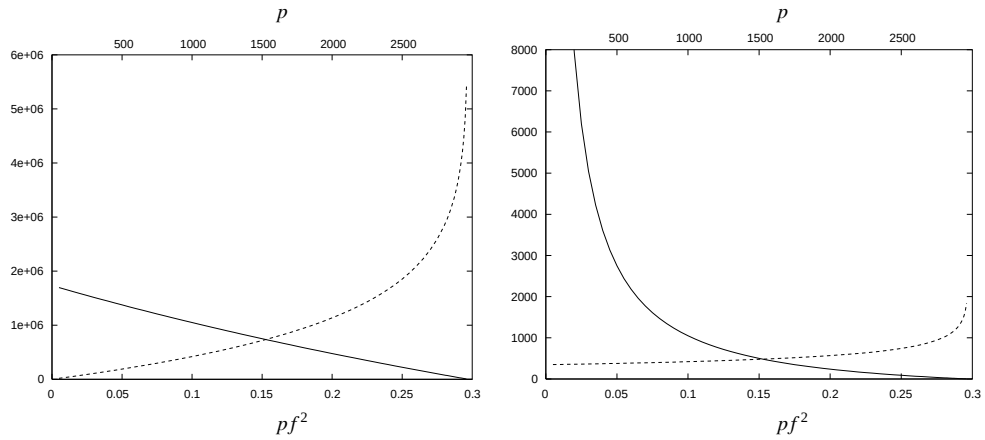

Figure 2. Number of presentations for learning (τ_+, dashed lines) and forgetting (τ, solid lines), as a function of the number of prototypes p (shown on top) and the rescaled number of prototypes $\alpha = pf^2$ (bottom), for $q = 0.002$, $f = 0.01$, and $\rho = 1$. Left: total number of presentations; Right: number of presentations per prototype.

about 35 000 presentations (350 per class) to learn a new class, while one still has to wait a large number of presentations (about 1.6×10^6) to forget it.

Figure 1 shows the behaviour of g and g_+ as a function of p, in both low- and high-loading regimes. Figure 2 shows the number of presentations needed to learn/forget as a function of p (shown on top) and α (bottom). The left-hand figure shows the total number of presentations while the right-hand one shows the number of presentations per class. When the number of classes is small learning is much faster than forgetting. As p increases, learning becomes slower and forgetting faster, so that at $p \sim 1500$ learning becomes slower than forgetting. This is due to the fact the ICP g_+ becomes closer to its 'critical' value $g + \Delta g_c$, so that as soon as the prototypes are not shown any more, its ICP decreases to its critical value in a short time. In fact, the forgetting time goes to 0 as we come close to the storage capacity.

5.2. Learning prototypes from class members

The results of the previous section are now generalized to the case of classes with finite extent ($x > 0$). At each presentation a member is extracted from a class chosen at random.

5.2.1. Low loading. We set $p = \alpha/f$, where α is a finite parameter. The average potentiation level is now given by equation (D14). Unlike in the case of pure prototypes, for any $x > 0$ no memory of the initial conditions survives, since the dependence on the initial distribution g_0 has disappeared. g varies from $g = x/(2\rho + x)$ when α goes to zero to $g = x(2 - x)/(2\rho + x(2 - x))$ when α goes to infinity. For any α, the average potentiation level gradually increases with the class extent x, from $g = 0$ when x goes to zero to $g = 1/(1 + 2\rho)$ at $x = 1$. This is due to the fact that synapses which are not 'intra-class' synapses (the overwhelming majority of synapses when loading is low) have a finite probability of seeing both pre- and postsynaptic neurons active. This probability increases with the class extent x.

The average IC potentiation level is in the limit $f \to 0$, for any $0 < x < 1$,

$$g_+ = 1.$$

At $x = 1$ there is an abrupt transition and g_+ becomes equal to g. This discontinuity in g_+ is obtained only when $f \to 0$. Thus, in this limit any positive correlation between class members and their prototype will lead to perfect learning of the prototype. The crossover between perfect learning ($g_+ = 1$) and no learning ($g_+ = g$) occurs for $x \sim 1 - \mathrm{O}(f)$.

High loading. We consider now the case $p = \alpha/f^2$. The average potentiation level is now given by equation (D15). g increases with the rescaled number of classes α from

$$g = \frac{x(2-x)}{2\rho + x(2-x)}$$

at $\alpha = 0$ to

$$g = \frac{1}{1+2\rho}$$

when α goes to infinity.

The average IC potentiation level is given by equation (D16). g_+ decreases again from $g_+ = 1$ at $\alpha = 0$ to

$$g_+ = \frac{1}{1+2\rho}$$

when α goes to infinity. The decrease is more abrupt when x increases. When $x = 1$ we have $g_+ = g = 1/(1+2\rho)$ as in the low-loading case.

These results are illustrated in figure 3 in which we show how g and g_+ vary as a function of the number of classes p, at different class extents x. As expected, the IC potentiation level decreases significantly when the class extent x increases. For example, applying again the criterion $g_+ - g > 0.5$ for prototype retrieval, and taking $f = 0.01$ and $\rho = 1$, we find that the maximal number of stored classes drops from about 3000 at $x = 0$ to about 400 at $x = 0.5$.

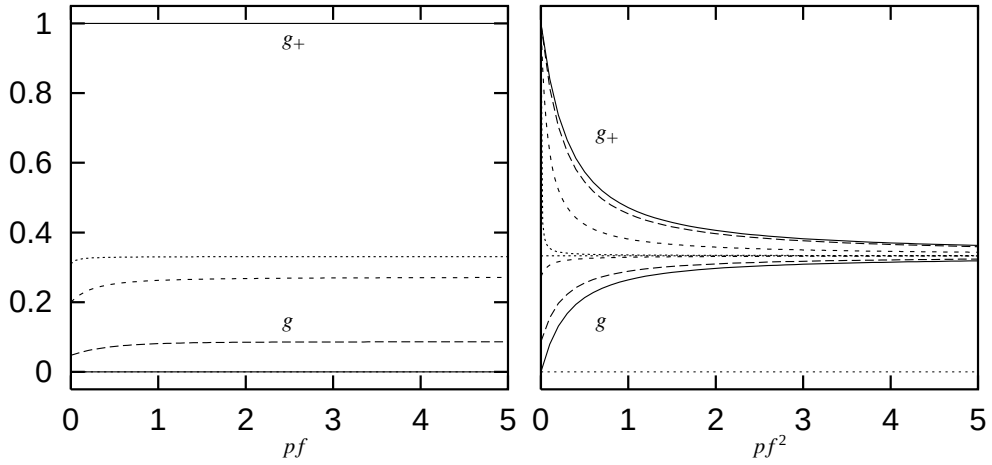

Figure 3. g_+ and g as a function of the number of classes p, at different levels of the class extent x: $x = 0$ (full lines); $x = 0.1$ (long dashed lines); $x = 0.5$ (short dashed lines); $x = 0.9$ (dotted lines). Left: low loading. Right: high loading.

5.3. Comparison between different types of LTD

We consider now the influence of varying the relative strength of homosynaptic and heterosynaptic LTD using the depression function of equation (6) instead of equation (5). We find that in the high-loading case the potentiation level is given by

$$g = \sum_{\Pi} \frac{\Pi}{\Pi + \alpha\rho(u+v)} \frac{\alpha^{\Pi} \exp(-\alpha)}{\Pi!}.$$

Analogously, the ICP and the corresponding expressions in the case of classes of finite extent can be obtained by replacing ρ by $\rho(u+v)/2$. Thus, the final expressions depend only on the average depression probability, which is proportional to $\rho(u+v)$, and homosynaptic and heterosynaptic depression or a combination of both types turn out to be completely equivalent. In the case of low loading, the potentiation level depends on the precise type of LTD, but the variation of g on u and v, at parity of $u+v$, is of the order of 0.001 and therefore smaller by several orders of magnitude than the difference between g_+ and g.

5.4. Prototypes versus class members

Next we ask whether the synaptic matrix is more correlated with the class prototypes or with the class members which have been shown to the network. To calculate more precisely the value of q at which the network is more correlated with classes than with prototypes, we return to equation (10) giving the probability distribution of a generic synaptic efficacy J_{ij} for an arbitrary sequence. Instead of averaging over all realizations of the sequence of presentations, we average over all realizations of sequences *that contain one specific instance $\mu\nu$ shown t presentations ago in the past*. After some algebra similar to that described in section 4, we find

$$\langle G_{ij} \rangle = \frac{\langle a_{ij} \rangle}{\langle a_{ij} + b_{ij} \rangle} + \langle \lambda_{ij} \rangle^t \left(a_{ij}(t) - \langle a_{ij} \rangle \frac{a_{ij}(t) + b_{ij}(t)}{\langle a_{ij} + b_{ij} \rangle} \right) \tag{31}$$

in which $a_{ij}(t)$ and $b_{ij}(t)$ are the terms due to presentation of instance $\mu\nu$.

We proceed now by considering the different types of synapse involved (to simplify things we consider $t = 0$, i.e. we look at the last shown instance, which is also the most correlated with the synaptic matrix).

- Synapses for which $\eta_i^{\mu}\eta_j^{\mu} = 1$, $\eta_i^{\mu\nu}\eta_j^{\mu\nu} = 1$ (both 'intra-class' and 'intra-example' synapses): for these synapses

  $$\frac{\langle a_{ij} \rangle}{\langle a_{ij} + b_{ij} \rangle} = g_+ \qquad a_{ij}(0) = q \qquad b_{ij}(0) = 0$$

  and thus

  $$\langle G_{ij} \rangle = g_+ + q(1 - g_+). \tag{32}$$

- Synapses for which $\eta_i^{\mu}\eta_j^{\mu} = 1$ but $\eta_i^{\mu\nu}\eta_j^{\mu\nu} = 0$ ('intra-class' but not 'intra-example'): for these synapses

  $$\frac{\langle a_{ij} \rangle}{\langle a_{ij} + b_{ij} \rangle} = g_+ \qquad a_{ij}(0) = 0 \qquad b_{ij}(0) = \mathrm{O}(f)$$

  and

  $$\langle G_{ij} \rangle = g_+ + \mathrm{O}(f). \tag{33}$$

- Synapses for which $\eta_i^\mu \eta_j^\mu = 0$, $\eta_i^{\mu\nu} \eta_j^{\mu\nu} = 1$ ('intra-example' but not 'intra-class'):

$$\frac{\langle a_{ij} \rangle}{\langle a_{ij} + b_{ij} \rangle} = g \qquad a_{ij}(0) = q \qquad b_{ij}(0) = 0$$

and

$$\langle G_{ij} \rangle = g + q(1 - g). \tag{34}$$

To determine whether the network is more correlated with the class prototype or with the last shown example we have to compare equations (33) and (34). If 'intra-class but not intra-example' synapses are on average stronger than 'intra-example but not intra-class', i.e. if

$$q < \frac{g_+ - g}{1 - g} \tag{35}$$

the correlation with the prototype will be stronger than the correlation with any example, otherwise the last example has been learned better than the class prototype. To determine how many of the last examples shown are more correlated than the prototype we should turn back to the t-dependent expression. Equation (35), in the low-loading case, simplifies to $q < 1$; thus, in this case, the synaptic matrix is necessarily more correlated with the prototype than with any example shown. In the high-loading regime, for any finite α and x, if q is low enough, as determined by the condition (35), in which g_+ and g are given by equations (D6) and (D7), the synaptic matrix will always be more correlated with prototypes of the classes, than with any of the examples it has seen.

Thus, in a situation of slow learning, the synaptic matrix is necessarily more correlated with the prototypes than with the class examples. The network categorizes input stimuli.

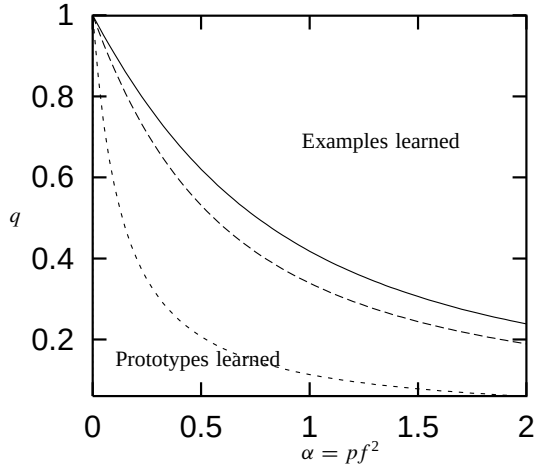

Figure 4. 'Critical' q below which class prototypes are learned better than shown examples, for different class extents: $x = 0$ (full lines); $x = 0.1$ (long dashed lines); $x = 0.5$ (short dashed lines).

These results are summarized in figure 4. This figure shows the regions in q–α plane in which prototypes or instances are learned. Note that as α or x become larger, the critical transition probability q for which prototypes are learned better than shown examples becomes lower. A similar result had been obtained in (Fusi 1995) in the case of a single class shown together with random stimuli.

6. Simulations

The learning dynamics used in the simulations is defined in section 2.3. We take $\rho = 1$, so that $q_+ = q$ and $q_- = fq$. The probability that a synapse is in its high state when the simulation starts is always set to zero.

6.1. Variability from sequence to sequence as a function of transition probability

We have shown in section 4.3 and appendix B that the variability of both potentiation level and intra-class potentiation level goes to zero as q tends to zero. In order to test the quantitative predictions of the theory we simulated the behaviour of a fully connected network of $N = 1000$ neurons. We generated $P = 15$ prototypes. The class amplitude x was set to zero. The network was presented S randomly chosen sequences. The length of these sequences was such that the network could reach the asymptotic regime, i.e. the number of presentations was chosen to be much larger than the forgetting time constants. At the end of each sequence we calculated the global connectivity ($\mathcal{G}$) and the ICP ($\mathcal{G}^\mu$) for each of the 15 classes:

$$\mathcal{G}_s = \frac{1}{N(N-1)} \sum_{i \neq j} J_{ij}^s \qquad \mathcal{G}_s^\mu = \frac{1}{N_+^\mu (N_+^\mu - 1)} \sum_{i \neq j} J_{ij}^s \xi_i^\mu \xi_j^\mu$$

where N_+^μ is the number of active neurons in the μth prototype and s is the index of the sequence.

At the end of the simulation we computed:

$$\Delta\mathcal{G}^2 = \frac{1}{S-1} \sum_{s=1}^{S} (\mathcal{G}_s - \langle\mathcal{G}\rangle)^2 \qquad (\Delta\mathcal{G}^\mu)^2 = \frac{1}{S-1} \sum_{s=1}^{S} (\mathcal{G}_s^\mu - \langle\mathcal{G}^\mu\rangle)^2$$

where $\langle\mathcal{G}\rangle = (1/S)\sum_{s=1}^{S} \mathcal{G}_s$ and $\langle\mathcal{G}^\mu\rangle = (1/S)\sum_{s=1}^{S} \mathcal{G}_s^\mu$. In figure 5 we compare the results of the simulations and the theoretical predictions at different values of the transition probability q. It shows that when q is small the theory is in good agreement with the simulation results.

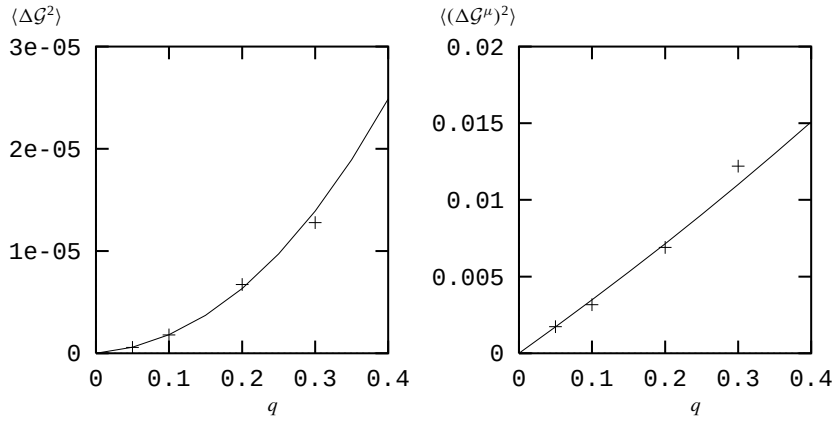

Figure 5. Variability of the global potentiation level (left) and of the ICP (right). Theoretical predictions (solid lines) and simulation results (+). The variability of the ICP is averaged over the 15 classes. The parameters are: $N = 1000$, $P = 15$, $f = 0.2$, $S = 50$.

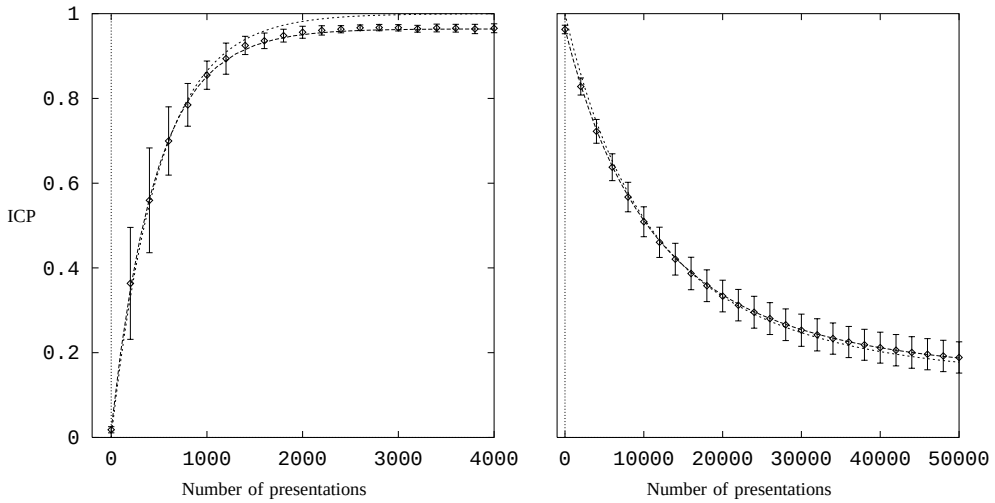

Figure 6. Learning and forgetting. Left: ICP of a prototype presented to the network to be learned versus the number of presentations. Right: ICP of the same prototype when it is not presented any more to the network (forgetting). Each diamond represents the mean over 16 simulations (error bars: standard deviation). Long dashed lines: theoretical prediction, equations (27) and (28), with the parameters of the simulations: $N = 3000$, $p = 50$, $f = 0.02$, $q = 0.1$. Short dashed lines: theoretical prediction in the sparse coding limit, equations (D3) and (D4). Note that for this value of f there is a small difference between the two theoretical predictions.

6.2. Learning and forgetting rate

6.2.1. Learning and forgetting pure prototypes. The behaviour of the ICP of a stimulus that is added or removed from the environment is described by the learning and forgetting functions defined in section 4.4.

A network of $N = 3000$ neurons was simulated, with a set of $P = 50$ uncorrelated prototypes with $f = 0.02$ (corresponding to $\alpha = 1$ in the low-loading regime, see section 5.1), and $x = 0$. The transition probability was set to $q = 0.1$. The network was presented a sequence of stimuli, long enough to forget the initial state of the synapses. Then, at the presentation labelled 0 on the horizontal axis of each plot in figure 6, one of the stimuli is removed from the set and a new one is added. The network is shown the sequence of stimuli randomly extracted from the new set. After each presentation we record the ICP of the prototype that has been removed and the ICP of the one that has been added.

The results of simulations are plotted in figure 6. It shows that the theoretical predictions for g, g_+, $\phi(t)$ and $\phi^\mu(t)$ derived in section 4 are in good agreement with the simulation results.

6.2.2. Learning and forgetting classes with finite extent. Class amplitude was then set to $x = 0.3$. At the beginning of the simulation we generated 50 uncorrelated patterns that will be used as the prototypes of 50 classes. During the first part of the simulation we randomly extract one of the prototypes at each time step and generate a pattern belonging to its class using the procedure of section 2.2. At presentation 0 in the plots of figure 7 one of the prototypes of the set is replaced by a new one. Then the new set is presented to the network and the two ICPs of the new prototype (learning curve) and of the old one

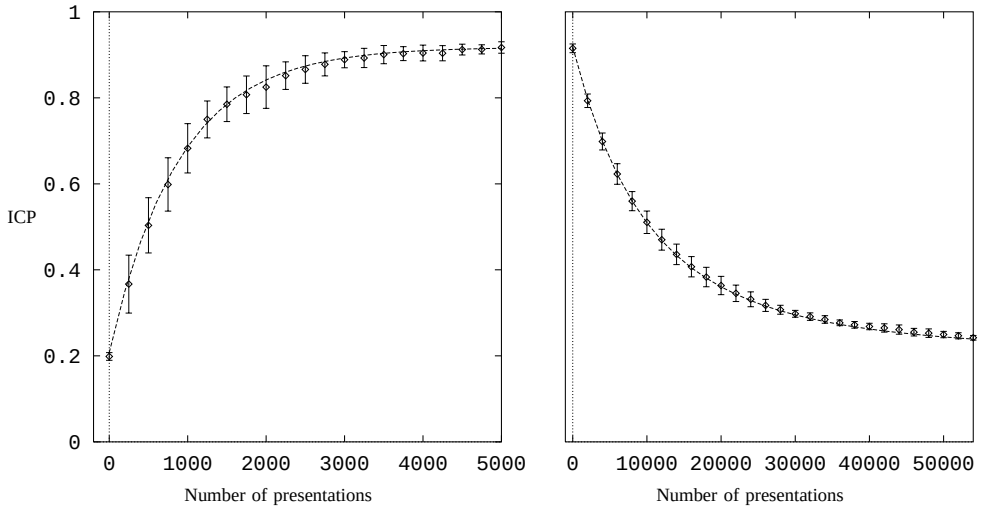

Figure 7. Learning (left) and forgetting (right) curves for classes of stimuli. The parameters used in the simulations are: $N = 3000$, $P = 50$, $f = 0.02$, $q = 0.1$, $x = 0.3$. Long dashed line: theory at finite f, equations (27) and (28). Short dashed line: theory in the limit $f = 0$, equations (D3) and (D4). The agreement with the theoretical predictions is even better than the case of pure prototypes. See the discussion in the text.

(forgetting curve) are estimated at each time step.

The results of simulations and the theoretical predictions are compared in figure 7. Note that in the case of the forgetting curve the introduction of classes reduces the variability of the ICP from simulation to simulation with respect to the pure prototype case. This can be explained by observing that the effect of showing a class is less dependent on the particular choice of prototypes. Thus, the quantities describing the behaviour of a network that learns classes of stimuli are better approximated by the population-averaged values.

7. Discussion

In this paper we have discussed a simple and biologically motivated learning dynamics, when classes of stimuli are shown repeatedly to the network. We have shown that in a slow learning scenario, the statistical properties of the synaptic matrix are essentially independent of the particular sequence of presentation. The learning dynamics we have considered in detail was motivated by experimental data on both LTP and LTD (Bliss and Collingridge 1993, Christie *et al* 1994). There is still some controversy about which patterns of activity provoke LTD. One interesting conclusion from the simple model presented in this paper is that with sparsely coded patterns and in the high-loading regime, the degree of correlation between the synaptic matrix and class prototypes shown to the network is independent on the precise form of LTD, provided the balance between LTP and LTD is kept fixed, i.e. globally the modifications induced by LTP and LTD are of the same order.

Simulations have shown that a neural network with analogue neurons, represented by a current-to-rate transduction function, is able to stabilize attractors correlated with shown prototypes, as soon as the correlation between these prototypes and the synaptic matrix is high enough (Amit and Brunel 1995). This is usually the critical parameter determining whether the network can function as an autoassociative memory (e.g. Willshaw *et al* 1969,

Nadal and Toulouse 1990). A detailed study of the neuronal dynamics and how it is influenced by synaptic dynamics, for various neuronal models will be presented in a separate publication.

In the stochastic learning formalism the speed of learning is mainly controlled by the LTP transition probability q. Different values of q will yield quite different behaviours: if q is high the network will learn in a single presentation individual examples that are shown to it, but will forget them, if they are not shown again, after a relatively short time. If instead q is low the network will need many presentations before it learns something, but conversely it will be able to 'extract' class prototypes from class members. To forget a class a large-scale change in the input statistics will be necessary. One might speculate that different areas in the brain might use these two different strategies. For example, it has been hypothesized that hippocampus serves as a short-, or intermediate-term memory storage (e.g. Marr 1971, Rolls 1990). In this case, it would be useful to have a strong probability of potentiation there, in order to learn new events in one shot. On the other hand, other areas, for example in neocortex, might need many repetitions of correlated stimuli before a prototype is extracted and stored in a longer-term memory, as experiments in IT cortex seem to suggest (Miyashita 1993). A learning process with a low potentiation transition probability, such as the one considered in this paper, would be ideally suited for that.

Acknowledgments

We thank Daniel Amit for many helpful discussions and for a critical reading of the manuscript that greatly improved a previous version. NB acknowledges a fellowship of Programme Cognisciences, CNRS, France, and the hospitality of INFN, Università di Roma, at the beginning of this work. The work was supported in part by a Human Mobility grant from the EC.

Appendix A. Synaptic distribution as a function of the sequence of stimuli

The probabilities of the synapse ij being in the excited state ($G_{ij}(T) \equiv \Pr(J_{ij} = 1, T)$) or background state ($1 - G_{ij}(T) \equiv \Pr(J_{ij} = 0, T)$) at time T can be written as a function of the transition matrices $M_{ij}(t)$ with $0 < t \leqslant T$ as

$$\begin{aligned}(G_{ij}(T), 1 - G_{ij}(T)) &= (G_{ij}(0), 1 - G_{ij}(0))\, M_{ij}(1)\, M_{ij}(2)\, \ldots\, M_{ij}(T) \\ &= (G_{ij}(0), 1 - G_{ij}(0))\, \mathcal{M}_{ij}(T) \end{aligned} \tag{A1}$$

where $\mathcal{M}_{ij}(T)$ is the transition matrix corresponding to the presentation of the T stimuli. We denote the product of the last τ matrices by $\mathcal{M}_{ij}(\tau)$:

$$\mathcal{M}_{ij}(\tau) = \begin{pmatrix} 1 - B_{ij}^{\tau} & B_{ij}^{\tau} \\ A_{ij}^{\tau} & 1 - A_{ij}^{\tau} \end{pmatrix} = M_{ij}(T - \tau + 1)\, M_{ij}(T - \tau + 2)\, \ldots\, M_{ij}(T).$$

The explicit expression for A and B can be deduced by solving the following recurrence relations:

$$A_{ij}^{\tau+1} = A_{ij}^{\tau} + a^{T-\tau}(1 - A_{ij}^{\tau} - B_{ij}^{\tau}) \tag{A2}$$

$$B_{ij}^{\tau+1} = B_{ij}^{\tau} + b^{T-\tau}(1 - A_{ij}^{\tau} - B_{ij}^{\tau}). \tag{A3}$$

First we compute the total probability that some transition is provoked by the last τ presentations, $C_{ij}^{\tau} = A_{ij}^{\tau} + B_{ij}^{\tau}$, by summing the two relations:

$$C_{ij}^{\tau} = 1 - \prod_{r=0}^{\tau-1} \lambda_{ij}^{T-r} = 1 - \prod_{s=T-\tau+1}^{T} \lambda_{ij}^{s}$$

where $\lambda_{ij}^{t} = 1 - a_{ij}^{t} - b_{ij}^{t}$. If we substitute C_{ij}^{τ} in equations (A2) and (A3) we find

$$A_{ij}^{T} = \sum_{t=1}^{T} a_{ij}^{t} \prod_{s=t+1}^{T} \lambda_{ij}^{s} \qquad B_{ij}^{T} = \sum_{t=1}^{T} b_{ij}^{t} \prod_{s=t+1}^{T} \lambda_{ij}^{s}$$

and using these relations together with equation (A1) we obtain:

$$P(J_{ij} = 1, T) = G_{ij}(T) = \sum_{t=1}^{T} a_{ij}^{t} \prod_{s=t+1}^{T} \lambda_{ij}^{s} + G_{ij}(0) \prod_{s=1}^{T} \lambda_{ij}^{s}.$$

Appendix B. Sequence variability of potentiation levels

The sequence to sequence variability of the asymptotic potentiation level is:

$$\Delta g^2 = \frac{1}{N^2(N-1)^2} \sum_{i \neq j, h \neq l} \langle \Delta G_{ij} \Delta G_{hl} \rangle. \tag{B1}$$

We first calculate $\langle G_{ij} G_{hl} \rangle$ for a single pair of synapses. In order to simplify the notation we write $a = a_{ij}$, $b = b_{ij}$, $\lambda = \lambda_{ij}$ to represent the variables of the synapse ij and $\alpha = a_{hl}$, $\beta = b_{hl}$, $\eta = \lambda_{hl}$ for the synapse hl. Using equation (10) this average becomes :

$$\langle G_{ij} G_{hl} \rangle = \left\langle \sum_{t,k=-\infty}^{T} a^t \alpha^k \prod_{s=t+1}^{T} \prod_{r=k+1}^{T} \lambda^s \eta^r \right\rangle = \langle A^2 + B + D \rangle$$

where the three terms of the RHS are:

$$A^2 = \sum_{t<T} a^t \alpha^t \prod_{s=t+1}^{T} \lambda^s \eta^s \qquad B = \sum_{t,k<t} a^t \alpha^k \prod_{s=t+1}^{T} \prod_{r=k+1}^{T} \lambda^s \eta^r$$

$$D = \sum_{k,t<k} a^t \alpha^k \prod_{s=t+1}^{T} \prod_{r=k+1}^{T} \lambda^s \eta^r.$$

Averaging over sequences we find:

$$\langle A^2 \rangle = \langle a\alpha \rangle \sum_{t<T} \langle \lambda\eta \rangle^{T-t} = \frac{\langle a\alpha \rangle}{1 - \langle \lambda\eta \rangle}$$

$$\langle B \rangle = \left\langle \sum_{t,k<t} a^t \eta^t \alpha^k \prod_{r=k+1}^{t-1} \eta^r \prod_{s=t+1}^{T} \lambda^s \eta^s \right\rangle = \langle \alpha \rangle \langle a\eta \rangle \sum_{t,k<t} \langle \eta \rangle^{t-k} \langle \lambda\eta \rangle^{T-t} = \frac{\langle \alpha \rangle \langle a\eta \rangle}{(1 - \langle \eta \rangle)(1 - \langle \lambda\eta \rangle)}$$

$$\langle D \rangle = \frac{\langle a \rangle \langle \alpha\lambda \rangle}{(1 - \langle \lambda \rangle)(1 - \langle \lambda\eta \rangle)}$$

and

$$\langle G_{ij} G_{hl} \rangle = \frac{1}{1 - \langle \lambda\eta \rangle} \left(\langle a\alpha \rangle + \frac{\langle \alpha \rangle \langle a\eta \rangle}{1 - \langle \eta \rangle} + \frac{\langle a \rangle \langle \alpha\lambda \rangle}{1 - \langle \lambda \rangle} \right). \tag{B2}$$

Using equation (12), we obtain:

$$\langle \Delta G_{ij} \Delta G_{hl} \rangle = \frac{\langle a\alpha \rangle \langle b \rangle \langle \beta \rangle + \langle b\beta \rangle \langle a \rangle \langle \alpha \rangle - \langle a\beta \rangle \langle \alpha \rangle \langle b \rangle - \langle \alpha b \rangle \langle a \rangle \langle \beta \rangle}{\langle a+b \rangle \langle \alpha+\beta \rangle (\langle a+b+\alpha+\beta \rangle - \langle (a+b)(\alpha+\beta) \rangle)}. \tag{B3}$$

The terms of the sum in expression (B1) can be divided into three classes: terms that have two pairs of equal indexes (e.g. $i = h,\ j = l$), $\langle \Delta G_{ij} \Delta G_{hl} \rangle = \langle \Delta G_{ij}^2 \rangle$; terms with one pair of equal indexes (e.g. $i = l,\ j \neq h$), $\langle \Delta G_{ij} \Delta G_{hl} \rangle = \langle \Delta G_{ij} \Delta G_{ih} \rangle$; and finally the class in which all the four indices are different. The expression (B1) becomes:

$$\Delta g^2 = \frac{1}{N^2(N-1)^2} \left[2 \sum_{i \neq j} \langle \Delta G_{ij}^2 \rangle + 4 \sum_{i \neq j \neq h} \langle \Delta G_{ij} \Delta G_{ih} \rangle + \sum_{i \neq j \neq h \neq k} \langle \Delta G_{ij} \Delta G_{hk} \rangle \right]. \quad \text{(B4)}$$

The variability of g^μ has the same form as the variability of g, the only difference being that the synapses of expression (8) see at least one potentiating prototype (the μth prototype). Thus, the average over populations must be calculated for $P - 1$ random prototypes and a potentiating one.

With the given dynamics equation (B3) can be written in the form

$$\langle \Delta G_{ij} \Delta G_{hl} \rangle \equiv q I_{ijhl} = q \frac{c_{ijhl}}{d_{ijhl} - q e_{ijhl}}$$

where:

$$c_{ijhl} = \chi^2 \left(\langle p_{ij} p_{hl} \rangle \langle d_{ij} \rangle \langle d_{hl} \rangle + \langle d_{ij} d_{hl} \rangle \langle p_{ij} \rangle \langle p_{hl} \rangle - \langle p_{ij} d_{hl} \rangle \langle p_{hl} \rangle \langle d_{ij} \rangle - \langle p_{hl} d_{ij} \rangle \langle p_{ij} \rangle \langle d_{hl} \rangle \right)$$
$$d_{ijhl} = \langle p_{ij} + \chi d_{ij} \rangle \langle p_{hl} + \chi d_{hl} \rangle \langle p_{ij} + \chi d_{ij} + p_{hl} + \chi d_{hl} \rangle$$
$$e_{ijhl} = \langle p_{ij} + \chi d_{ij} \rangle \langle p_{hl} + \chi d_{hl} \rangle \langle (p_{ij} + \chi d_{ij})(p_{hl} + \chi d_{hl}) \rangle$$

in which c, d and e are independent of q and $e < d$. The expression for p_{ij}, d_{ij} is given by equations (4) and (5). Thus, $\langle \Delta G_{ij} \Delta G_{hl} \rangle = \mathrm{O}(q)$ when q goes to zero, and consequently the fluctuations of both g and g^μ go to zero as q.

Appendix C. Potentiating and depressing distributions in the sparse coding limit

The joint distribution of the numbers of potentiating and depressing prototypes is

$$\Pr(\Pi, \Delta) = \frac{p!}{\Pi! \Delta! (p - \Pi - \Delta)!} f^{2\Pi} (2f(1-f))^{\Delta} (1-f)^{2(p-\Pi-\Delta)}$$
$$= \frac{1}{\Pi! \Delta!} f^{2\Pi} (2f(1-f))^{\Delta} A$$

where

$$A = \frac{p!}{(p - \Pi - \Delta)!} (1-f)^{2(p-\Pi-\Delta)}.$$

In the limit $p \to \infty$ we can apply Stirling's formula and obtain

$$A = \exp \Bigg[2(p - \Pi - \Delta) \ln(1-f) - p + \left(p + \frac{1}{2} \right) \ln p$$
$$+ p - \Pi - \Delta - \left(p - \Pi - \Delta - \frac{1}{2} \right) \ln(p - \Pi - \Delta) \Bigg].$$

We define new variables x, y and z by $\Pi = p f^2 x$, $\Delta = 2p(1-f) f y$, $z = f^2 x + 2f(1-f) y$, and obtain

$$A = p^{\Pi + \Delta} \exp \left[2p \ln(1-f)(1-z) - pz - p(1-z) \ln(1-z) - \frac{1}{2} \ln(1-z) \right]$$
$$A = p^{\Pi + \Delta} B.$$

Now we take the limit as $f \to 0$, neglecting all terms of order z, pz^3 and pf^3,

$$B = \exp\left[-2pf + 2pfz - pf^2 - \frac{1}{2}pz^2\right] = \exp\left[-2pf + pf^2 - 2pf^2(1-y)^2\right].$$

Putting everything together we obtain

$$\Pr(\Pi, \Delta) = \frac{(pf^2)^\Pi}{\Pi!}\exp(-pf^2)\frac{(2f(1-f))^\Delta}{\Delta!}\exp(-2pf(1-f)) \times \exp\left[-2pf^2\left(1 - \frac{\Delta}{2pf(1-f)}\right)^2\right]$$

plus terms of order f or pf^3. Thus, the joint distribution for Π and Δ decouples and becomes the product of a Poisson distribution with mean pf^2 for Π, multiplied by a distribution for Δ which is slightly different from a Poisson distribution. However, the last term gives a negligible contribution in both cases, $p \sim 1/f$ and $p \sim 1/f^2$, since in the latter case the variable $\Delta/[2pf(1-f)]$ has mean 1 and variance of order $1/pf \sim f$. Thus, in both cases the distribution of Δ is Poisson with mean $2pf(1-f)$. The 'intra-class' distribution can be calculated in the same way. The leading-order term is again the product of the two Poisson distributions.

Appendix D. Results in the sparse coding limit

Results are summarized in table 1.

Appendix D.1. Pure prototypes

We first study the case in which only class prototypes are shown, and transitions occur as defined by equations (4) and (5). Substituting (30) in (15), we can rewrite the sequence-averaged probability that synapse J_{ij} is potentiated, g_{ij}, as

$$g_{ij} = \frac{P_{ij}}{P_{ij} + f\rho D_{ij}}$$

and $\langle\lambda_{ij}\rangle$, as

$$\langle\lambda_{ij}\rangle = 1 - \frac{q}{p}\left(P_{ij} + f\rho D_{ij}\right).$$

Thus, from equations (18), (19), (27) and (28) we deduce the average potentiation level,

$$g = \sum_{\Pi,\Delta}\psi_{\mathrm{p}}(\Pi)\psi_{\mathrm{d}}(\Delta)\frac{\Pi}{\Pi + f\rho\Delta} \tag{D1}$$

the 'intra-class' potentiation level,

$$g_+ = \sum_{\Pi,\Delta}\psi_{\mathrm{p}}(\Pi)\psi_{\mathrm{d}}(\Delta)\frac{1+\Pi}{1+\Pi + f\rho\Delta} \tag{D2}$$

the forgetting function,

$$\phi(T) = g + \sum_{\Pi,\Delta}\psi_{\mathrm{p}}(\Pi)\psi_{\mathrm{d}}(\Delta)\left(g_+ - \frac{\Pi}{\Pi + f\rho\Delta}\right)\left(1 - \frac{q}{p}(\Pi + f\rho\Delta)\right)^T \tag{D3}$$

Table A1. Summary of results in the sparse coding limit.

Pure prototypes ($x = 0$), low loading
$g = g_0 \exp(-2\alpha)$
$g_+ = 1$
$\phi(T) = \exp\left[-2\alpha\left(1 - \exp\left[-\frac{qf^2\rho T}{\alpha}\right]\right)\right]$
$\phi_+(T) = 1 + (g-1)\exp\left(-\frac{qfT}{\alpha}\right)$
Pure prototypes ($x = 0$), high loading
$g = \sum_{\Pi} \frac{\Pi}{\Pi + 2\alpha\rho} \frac{\alpha^{\Pi}\exp(-\alpha)}{\Pi!}$
$g_+ = \sum_{\Pi=0} \psi_{\mathrm{p}}(\Pi) \frac{\Pi+1}{\Pi+1+2\alpha\rho}$
$\phi(T) = g + \exp\left(-2\rho q f^2 T\right) \sum_{\Pi} \left(g_+ - \frac{\Pi}{\Pi+2\alpha\rho}\right) \exp\left(-\frac{qf^2\Pi T}{\alpha}\right) \psi_{\mathrm{p}}(\Pi)$
$\phi_+(T) = \exp\left[-\left(2\rho + \frac{1}{\alpha}\right) qf^2 T\right] \sum_{\Pi} \left(\frac{\Pi+1}{\Pi+1+2\alpha\rho} - g\right) \exp\left(-\frac{qf^2\Pi T}{\alpha}\right) \psi_{\mathrm{p}}(\Pi)$
Classes of extent $x > 0$, low loading
$g = \sum_{\Delta} \psi_{\mathrm{d}}(\Delta) \frac{x(1-x)\Delta + \alpha x^2}{x(1-x)\Delta + \rho(1-x)\Delta + \alpha x(x+2\rho)}$
$g_+ = 1$
Classes of extent $x > 0$, high loading
$g = \sum_{\Pi} \frac{(1-x)^2\Pi + \alpha x(2-x)}{(1-x)^2\Pi + \alpha(2\rho + x(2-x))} \frac{\alpha^{\Pi}\exp(-\alpha)}{\Pi!}$
$g_+ = \sum_{\Pi} \frac{(1-x)^2(\Pi+1) + \alpha x(2-x)}{(1-x)^2(\Pi+1) + \alpha(2\rho + x(2-x))} \frac{\alpha^{\Pi}\exp(-\alpha)}{\Pi!}$

and the learning function

$$\phi_+(T) = g_+ + \sum_{\Pi,\Delta} \psi_{\mathrm{p}}(\Pi)\psi_{\mathrm{d}}(\Delta) \left(g - \frac{1+\Pi}{1+\Pi+f\rho\Delta}\right) \left(1 - \frac{q}{p}(1+\Pi+f\rho\Delta)\right)^T. \quad \text{(D4)}$$

In the low-loading case, $p = \alpha/f$, where α is a finite parameter, and $f \to 0$. To calculate the average potentiation level we first note that in the limit $f \to 0$, $\psi_{\mathrm{p}}(\Pi > 0) = 0$. A fraction $\psi_{\mathrm{d}}(\Delta = 0) = \exp(-2\alpha)$ of synapses sees no depressing prototype and therefore never experiences transitions; these synapses stay at their initial value. The remaining synapses $\psi_{\mathrm{d}}(\Delta > 0) = 1 - \exp(-2\alpha)$ see at least one depressing prototype and consequently will eventually be in their low state. Thus, the average potentiation level is simply

$$g = g_0 \exp(-2\alpha)$$

where g_0 is the initial potentiation level.

To calculate the average IC potentiation level, we have to consider synapses that see at least one potentiating event (the one corresponding to the considered class). The probability of seeing another potentiating prototype is negligible in the limit $f \to 0$, and thus equation (D2) becomes

$$g_+ = \sum_{\Delta} \frac{1}{1+f\rho\Delta} \psi_{\mathrm{d}}(\Delta) \quad \text{(D5)}$$

where ψ_d is the Poisson distribution with mean 2α, equation (29). In this case, Δ is finite in the terms that contribute to the sum in the RHS of equation (D5), so that $g_+ = 1 - O(f)$.

To calculate the forgetting function, we insert equation (29) in equation (D3), and obtain:

$$\phi(T) = \exp\left[-2\alpha\left(1 - \exp\left[-\frac{qf^2\rho T}{\alpha}\right]\right)\right].$$

For T small compared with $\alpha/(q\rho f^2)$ we have

$$\phi(T) \sim \exp(-2qf^2\rho T).$$

With a similar calculation it is easy to derive the learning function from equation (D4)

$$\phi_+(T) = 1 + (g-1)\exp\left(-\frac{qfT}{\alpha}\right).$$

We now consider the high-loading case, $p = \alpha/f^2$. In this case each synapse typically sees a few potentiating events, as described by the Poisson distribution, equation (29), and a very large number of depressing events, of the order of $1/f$. The distribution of $f\Delta$ becomes sharply peaked around 2α, with a variance of the order of $\sqrt{f}$. Thus, in the limit $f \to 0$ we can replace $f\Delta$ by its mean 2α.

Using equations (29) and (D1), we find that the average potentiation level is now given by

$$g = \sum_{\Pi} \frac{\Pi}{\Pi + 2\alpha\rho} \frac{\alpha^{\Pi}\exp(-\alpha)}{\Pi!}. \tag{D6}$$

When α goes to infinity, g goes to

$$g = \frac{1}{1+2\rho}.$$

The average IC potentiation level is

$$g_+ = \sum_{\Pi=0} \psi_p(\Pi)\frac{\Pi+1}{\Pi+1+2\alpha\rho} \tag{D7}$$

and after some algebra we find that it is simply related to g by the expression

$$g_+ = 1 - 2\rho g.$$

To calculate the forgetting function we use equation (D3), replace $f\Delta$ by its mean 2α and insert ψ_p of equation (29), and finally obtain

$$\phi(T) = g + \exp\left(-2\rho q f^2 T\right)\sum_{\Pi}\left(g_+ - \frac{\Pi}{\Pi+2\alpha\rho}\right)\exp\left(-\frac{qf^2\Pi T}{\alpha}\right)\psi_p(\Pi). \tag{D8}$$

The learning function can be obtained in a similar way from equation (D4)

$$\phi_+(T) = \exp\left[-\left(2\rho + \frac{1}{\alpha}\right)qf^2T\right]\sum_{\Pi}\left(\frac{\Pi+1}{\Pi+1+2\alpha\rho} - g\right)\exp\left(-\frac{qf^2\Pi T}{\alpha}\right)\psi_p(\Pi). \tag{D9}$$

Appendix D.2. Learning prototypes from class members

To calculate the statistical properties of the synaptic matrix we have to come back to equation (12). In this equation to perform the average over sequences we use equations (2) and (3) defining the distribution of examples: we find that the sequence-averaged probability for synapse ij to be in its high state is

$$g_{ij} = \frac{\tilde{P}_{ij}}{\tilde{P}_{ij} + f\rho\tilde{D}_{ij}}$$

where

$$\tilde{P}_{ij} = [1 - x(1-f)]^2 P_{ij} + fx[1 - x(1-f)]D_{ij} + (fx)^2(1 - P_{ij} - D_{ij}) \tag{D10}$$

$$\tilde{D}_{ij} = 2(1-f)x[1 - x(1-f)]P_{ij} + [1 - x + 2f(1-f)x^2]D_{ij} + 2fx(1-fx)(1 - P_{ij} - D_{ij}) \tag{D11}$$

in which P_{ij} (D_{ij}) are the usual numbers of potentiating (depressing) prototypes, as defined in equation (14).

When $x = 1$, i.e. class members are uncorrelated with the class prototypes, we have $P_{ij} = f^2$ and $D_{ij} = 2f(1-f)$, and thus the synaptic distribution becomes independent of the prototypes, as expected.

In the sparse coding limit we can again calculate the average potentiation levels, as a function of x. Using equations (D10) and (D11) and keeping only the dominant terms in the limit $f \to 0$, we find

$$g = \sum_{\Pi,\Delta} \psi_{\mathrm{p}}(\Pi)\psi_{\mathrm{d}}(\Delta)\frac{(1-x)^2\Pi + fx(1-x)\Delta + f^2x^2}{(1-x)^2\Pi + f(1-x)(x+\rho)\Delta + f^2x(x+2\rho)} \tag{D12}$$

$$g_+ = \sum_{\Pi,\Delta} \psi_{\mathrm{p}}(\Pi)\psi_{\mathrm{d}}(\Delta)\frac{(1-x)^2(1+\Pi) + fx(1-x)\Delta + f^2x^2}{(1-x)^2(1+\Pi) + f(1-x)(x+\rho)\Delta + f^2x(x+2\rho)}. \tag{D13}$$

We first set $p = \alpha/f$, were α is a finite parameter. As in the case of pure prototypes, the potentiation levels can be obtained by setting $\Pi = 0$ in equations (D12) and (D13). The average potentiation level is now

$$g = \sum_{\Delta} \psi_{\mathrm{d}}(\Delta)\frac{x(1-x)\Delta + \alpha x^2}{x(1-x)\Delta + \rho(1-x)\Delta + \alpha x(x+2\rho)}. \tag{D14}$$

The average IC potentiation level is obtained using equations (29) and (D13), and we find that in the limit $f \to 0$, for any $0 < x < 1$,

$$g_+ = 1.$$

We now consider the case $p = \alpha/f^2$. We again use equations (D12) and (D13), in which we set $\Delta = 2\alpha$, as in the case for $x = 0$. The average potentiation level is now

$$g = \sum_{\Pi} \frac{(1-x)^2\Pi + \alpha x(2-x)}{(1-x)^2\Pi + \alpha(2\rho + x(2-x))}\frac{\alpha^{\Pi}\exp(-\alpha)}{\Pi!}. \tag{D15}$$

The average IC potentiation level is

$$g_+ = \sum_{\Pi} \frac{(1-x)^2(\Pi+1) + \alpha x(2-x)}{(1-x)^2(\Pi+1) + \alpha(2\rho + x(2-x))}\frac{\alpha^{\Pi}\exp(-\alpha)}{\Pi!}. \tag{D16}$$

References

Amit D J 1989 *Modeling Brain Function* (New York: Cambridge University Press)

Amit D J and Brunel N 1995 Learning internal representations in an attractor neural network with analogue neurons *Network: Comput. Neural Syst.* **6** 359

——1997 Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex *Cerebral Cortex* **7** 237

Amit D J, Brunel N and Tsodyks M V 1994 Correlations of cortical Hebbian reverberations: experiment versus theory *J. Neurosci*. **14** 6435

Amit D J and Fusi S 1992 Constraints on learning in dynamic synapses *Network: Comput. Neural Syst.* **3** 443

——1994 Dynamic learning in neural networks with material synapses *Neural Comput*. **6** 957

Amit D J, Fusi S and Yakovlev V 1997 A paradigmatic attractor cell in IT *Neural Comput.* **9** 1101

Annunziato M 1995 Hardware implementation of an attractor neural network with integrate-and-fire neurons and stochastic learning *Thesis* Università di Roma 'La Sapienza' (in Italian)

Artola A and Singer W 1993 Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation *Trends Neurosci.* **16** 480

Badoni D, Bertazzoni S, Buglioni S, Salina G, Amit D J and Fusi S 1995 Electronic implementation of an analogue neural network with stochastic transitions *Network: Comput. Neural Syst.* **6** 125

Bienenstock E, Cooper L and Munro 1982 Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex *J. Neurosci*. **2** 32

Bliss T V P and Collingridge G L 1993 A synaptic model of memory: long-term potentiation in the hippocampus *Nature* **361** 31

Brunel N 1994 Dynamics of an attractor neural network converting temporal into spatial correlations *Network: Comput. Neural Syst.* **5** 449

——Brunel N 1996 Hebbian learning of context in recurrent neural networks *Neural Comput*. **8** 1677

Christie B R, Kerr D S and Abraham W C 1994 Flip side of synaptic plasticity: long-term depression mechanisms in the hippocampus *Hippocampus* **4** 127

Fusi S 1995 Prototype extraction in material attractor neural networks with stochastic dynamic learning *Proc. SPIE 2492 (Proc. SPIE'95, Applications and Science of Artificial Neural Networks, Orlando, FL)* ed S K Rogers and D W Ruck, p 1027

Fuster J 1995 *Memory in the Cerebral Cortex* (Cambridge, MA: MIT Press)

Hebb D O 1949 *The Organization of Behavior: a Neuropsychological Theory* (New York: Wiley)

Heskes T M and Kappen B 1991 Learning processes in neural networks *Phys. Rev.* A **44** 2718

Hopfield J J 1982 Neural networks and systems with emergent selective computational abilities *Proc. Natl Acad. Sci. USA* **79** 2554

Marr D 1971 Simple memory: a theory for archicortex *Phil. Trans. R. Soc.* B **262** 21

Meunier C and Nadal J P 1995 Sparsely coded neural networks *Handbook of Brain Theory and Neural Networks* ed M Arbib (Cambridge, MA: MIT Press) p 899

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

——1993 Inferior temporal cortex: where visual perception meets memory *Ann. Rev. Neurosci*. **16** 245

Nadal J P and Toulouse G 1990 *Network: Comput. Neural Syst.* **1** 61

Nadal J P, Toulouse G, Changeux J P and Dehaene S 1986 Networks of formal neurons and memory palimpsests *Europhys. Lett.* **1** 535

Parisi G 1986 A memory which forgets *J. Phys. A: Math. Gen.* **19** L617

Rolls E T 1990 Principles underlying the representation and storage of information in neuronal networks in the primate hippocampus and cerebral cortex *An Introduction to Neural and Electronic Networks* ed Zornetzer *et al* (San Diego: Academic)

Sejnowski T J 1977 Storing covariance with nonlinearly interacting neurons *J. Math. Biol*. **4** 303

Shinomoto S 1987 Memory maintenance in neural networks *J. Phys. A: Math. Gen.* **20** L1305

Tsodyks M V 1990 Associative memory in neural networks with binary synapses *Mod. Phys. Lett.* B **4** 713

Tsodyks M V and Feigel'man M V 1988 The enhanced storage capacity in neural networks with low activity level *Europhys. Lett.* **46** 101

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

Wong K Y M, Kahn P E and Sherrington D 1991 A neural network model of working memory exhibiting primacy and recency *J. Phys. A: Math. Gen.* **24** 1119