Communicated by David Willshaw

Learning in Neural Networks with Material Synapses

Daniel J. Amit*
INFN, Sezione di Roma, Istituto di Fisica, Università di Roma, La Sapienza, P.le Aldo Moro, Roma, Italy

Stefano Fusi
INFN, Sezione Sanità, Viale Regina Elena, 299, Roma, Italy

We discuss the long term maintenance of acquired memory in synaptic connections of a perpetually learning electronic device. This is affected by ascribing each synapse a finite number of stable states in which it can maintain for indefinitely long periods. Learning uncorrelated stimuli is expressed as a stochastic process produced by the neural activities on the synapses. In several interesting cases the stochastic process can be analyzed in detail, leading to a clarification of the performance of the network, as an associative memory, during the process of uninterrupted learning. The stochastic nature of the process and the existence of an asymptotic distribution for the synaptic values in the network imply generically that the memory is a palimpsest but capacity is as low as $\log N$ for a network of N neurons. The only way we find for avoiding this tight constraint is to allow the parameters governing the learning process (the coding level of the stimuli; the transition probabilities for potentiation and depression and the number of stable synaptic levels) to depend on the number of neurons. It is shown that a network with synapses that have two stable states can dynamically learn with optimal storage efficiency, be a palimpsest, and maintain its (associative) memory for an indefinitely long time provided the coding level is low and depression is equilibrated against potentiation. We suggest that an option so easily implementable in material devices would not have been overlooked by biology. Finally we discuss the stochastic learning on synapses with variable number of stable synaptic states.

1 Introduction

1.1 Memory Maintenance on Long Time Scales. A material neural network that is supposed to learn dynamically receives an uninterrupted flow of *uncorrelated* stimuli to be learned. The stimuli impinge on neural elements connected by synapses. An incoming stimulus imposes a certain activity distribution on the neural elements and each pair of neurons

*On leave of absence from Racah Institute of Physics.

generates a source for the learning by the synapse connecting them. On a short time scale it may be reasonable to assume that a synapse can modify its efficacy in an analog way, as would be the case for a capacitor. On long time scales, if memory is to be maintained even in the absence of stimuli and of neural activity, it is more likely that a synapse can preserve only a relatively small set of stable values. These we would identify with LTP. For the capacitor this is implemented by an asynchronous, continuous, stochastic threshold controlled refresh mechanism (Amit *et al.* 1992; Badoni *et al.* 1992). The discretized long-term synaptic values achieved this way must allow the network to act as an associative memory.

1.2 Learning as a Stochastic Process and Palimpsest Memory. Learning is a stochastic process either due to the nature of the data or due to the dynamics of synaptic modification. A stimulus in the sequence presented to the network is represented by a set of activity levels imposed on the neurons during its presentation. Since the stimuli are assumed uncorrelated each synapse will see a random sequence of pairs of activities on the two neurons connected by it. This is one source of stochasticity. We denote the information arriving on a given neuron by a binary variable ξ, indicating whether the corresponding neuron does or does not carry information. The second source of stochasticity is due to two possible factors. The actual coding of information on the neurons may be analog and hence the effect on the synapse may not be the same when the presented pattern has information represented with different amplitudes (such as different spike rates). Moreover, even given the same incoming pair of neural activities, it may still be the case that the transition from one stable synaptic state to another may not be deterministic (there may be noise in the threshold for the synaptic transition from one stable synaptic state to another). In other words, even upon the arrival of the same pair of information coding discrete variables a synapse will undergo the implied transition with probability that may be lower than unity.

As a consequence the presentation of a sequence of uncorrelated stimuli induces a Markovian process on the set of values of the $N(N-1)$ synapses. More formally, the probability that a synapse makes a transition $J \to J'$ is a product of $p_1(\xi, \tilde{\xi})$, the probability of the arrival of the pair $\xi, \tilde{\xi}$ on the two neurons connected by the synapse, and the probability that given that pair the transition takes place, $p_2(J \to J' \mid \xi, \tilde{\xi})$. We shall further assume that a given pair $\xi, \tilde{\xi}$ can produce a transition between a single pair of neighboring synaptic states, or no transition at all.

The resulting Markovian process is a walk on the *finite* set of stable synaptic values and will be described by the probability distribution function of the synaptic values. In particular, the conditional distribution function $\rho_J^p(\xi, \tilde{\xi})$ of obtaining the value J following the presentation

of p patterns the first of which imposed $\xi, \tilde{\xi}$ on the synapse satisfies the evolution equation:

$$\rho_J^p(\xi, \tilde{\xi}) = \sum_K \rho_K^1(\xi, \tilde{\xi})(M^{p-1})_{KJ} \tag{1.1}$$

in which M_{KJ} is the transition matrix whose elements are determined by the probabilities discussed above and the index J runs over all the stable synaptic states.

The first conclusion is that this type of dynamics is generically ergodic (see, e.g., Section 3). When the number of presented patterns becomes very large

$$\rho_J^p(\xi, \tilde{\xi}) \to \rho_J^\infty$$

which is independent of $\xi, \tilde{\xi}$. This makes a memory of this type a *palimpsest* (Nadal *et al.* 1986). In other words, patterns learned far in the past are erased by new patterns learned subsequently in sharp contrast to memories of the Hopfield or the Willshaw types (Hopfield 1982; Willshaw 1969). In the latter, when considered as a learning dynamics, following a large number of presentations all memory is destroyed (Amit *et al.* 1987).

An immediate implication of the existence of an asymptotic distribution of synaptic values, for a network that is to be available for learning for indefinitely long periods, is that the generic initial distribution on top of which learning new patterns is to take place is the asymptotic distribution ρ_J^∞. Having an asymptotic distribution is a necessary condition for palimpsest behavior. It is not sufficient. The asymptotic distribution must be such as to allow the learning process to imprint new stimuli upon it. A counterexample is provided by the Willshaw (1969) model, in which the asymptotic distribution is a synaptic distribution for which all synapses have the value +1 with probability 1. The presentation of any subsequent pattern will leave this distribution invariant and no retrieval is possible (see also Section 5).

To have a functioning learning network with a finite number of synaptic states, the presentation of a *given* new stimulus must change the conditional distribution $\rho_J(\xi, \tilde{\xi})$. Following the presentation of a given pattern consecutive presentations drive the conditional distribution back toward the asymptotic form, making the effect of the initial pattern progressively weaker. The question of the number of patterns that can be retrieved reduces therefore to the question about the age (distance into the past) of the oldest pattern that can still be retrieved, despite the effect of the subsequent patterns. Given the palimpsestic nature of the process, younger patterns can be retrieved a fortiori.

1.3 The Findings. We analyze the learning process as described above for a wide variety of cases. One main conclusion, already noticed in

Amit and Fusi (1992), is that if all parameters such as number of states per synapse; coding level in stimuli, and transition probabilities of a synapse for a given pair of neural activity variables, are independent of the number of neurons in the network, then at most $\log N$ patterns can be retrieved.

Making some of these parameters N-dependent one can do better. If the number of synaptic states increases with N, as fast as $\sqrt{N}$, then one can reach a storage of order N. This was also observed in Nadal *et al.* (1986) and Parisi (1986). Going beyond $\sqrt{N}$ in the number of states destroys the palimpsest behavior. Special initial synaptic conditions become required and the network suffers from the *blackout* effect, that is, all memories disappear together.

We then study a network with two states per synapse. In this case we find that if the coding level in the arriving stimuli is as low as $\log N/N$ one can reach storage capacities as high as $N/\log N$. For this type of patterns it was found (Willshaw 1969) that a network with two state synapses could have the optimal storage of $N^2/\log^2 N$ patterns. This is the price paid here for uninterrupted learning. Yet, when the intrinsic synaptic transition probabilities compensate for the coding level to make the mean number of up transitions (potentiation) of the same order as the number of down transitions (depression), one recovers optimal storage and enjoys continual learning. This additional requirement finds an interesting echo in recent experiments on potentiation and depression in hippocampal slices (Stanton and Sejnowski 1989).

2 Criteria for Retrieval

In the simple case of autoassociative memory the possibility of retrieving a memory is determined by the distribution of depolarizations among the neurons in the network upon the presentation of one of the previously memorized patterns. If that distribution is such that there exists a threshold that separates the depolarization of neurons that had been active in the learned pattern from those that had been quiescent, retrieval is in principle possible. The situation is even better if one can show that the relevant threshold can be plausibly generated by the neural dynamics. Retrieval is impossible, without errors if the two distributions of depolarizations overlap significantly (see, e.g., Weisbuch and Fogelman-Soulié 1985). The distribution of postsynaptic inputs is determined in turn by the collection of synaptic values.

The conditional distribution, equation 1.1, allows for the computation of the (conditional) mean of the synaptic input to a neuron that p patterns into the past had activity ξ. Similarly, we can compute the fluctuations of the postsynaptic input. If the sequence of afferent stimuli to be learned is $\xi_i^1, \xi_i^2 \ldots \xi_i^p$, then the synaptic input to neuron i upon presentation of the oldest memory ξ^1, following the learning of the entire sequence is

$$h_i^p = \frac{1}{N}\sum_{j=1}^{N} J_{ij}(p)\xi_j^1$$

where $J_{ij}(p)$ is the synaptic efficacy following the learning of the p patterns, and $1/N$ is a normalizing constant introduced for convenience. The synaptic inputs, h_i^p, can be classified according to the value that was imposed on neuron i during the imposition of pattern number 1. If the neural activity is coded by a binary variable ($\xi_i = \zeta_1, \zeta_2$), there will be two distributions of synaptic inputs: one for neurons that saw the value ζ_1 when ξ^1 was presented and another for those that saw ζ_2. The values of the input in each class have a conditional mean:

$$\langle h_i^p \rangle_\xi = \sum_{\tilde{\xi}} \rho_{\tilde{\xi}} \sum_J J\tilde{\xi}\rho_J^p(\xi, \tilde{\xi}) \tag{2.1}$$

where the conditional expectation $\langle \cdots \rangle_\xi$ is defined as

$$\langle f \rangle_\xi = \mathrm{E}(f \mid \xi_i^1 = \xi)$$

and ρ_ξ is the probability that a neuron had activity ξ when the network was stimulated by ξ^1. The expectation is over all the ξ_j^μ with $\mu > 1$ and $j = 1, \ldots, N$, and on ξ_j^1 with j different from i. In other words, this is the mean input to a neuron in the population that had activity ξ upon the presentation of ξ^1. The signal, for the binary case can be defined as

$$S = \langle h_i^p \rangle_{\zeta_1} - \langle h_i^p \rangle_{\zeta_2} \tag{2.2}$$

If $|S|$ is significantly greater than the sum of the noises around each of the mean h_i^p for the two values of ξ, then a threshold can be found that will separate correctly the two outcomes ζ_1,ζ_2 to reproduce the retrieved pattern. S can be written in terms of the conditional probabilities as

$$S = \sum_{\tilde{\xi}} \rho_{\tilde{\xi}} \sum_J J\tilde{\xi}[\rho_J^p(\zeta_1, \tilde{\xi}) - \rho_J^p(\zeta_2, \tilde{\xi})] \tag{2.3}$$

where the sum on $\tilde{\xi}$ extends over the two possible values of the activity ξ_j^1 and J runs over all n values of the stable synaptic states.

The fluctuations of the two h_i^p are estimated by

$$R^2(\xi) = \langle (h_i^p - \langle h_i^p \rangle_\xi)^2 \rangle_\xi$$

If the random variables h_i^p are gaussian, then total noise is

$$R^2 = \frac{1}{2}[R^2(\zeta_1) + R^2(\zeta_2)]$$

The computation of each of the current variances is complicated by the fact that it involves means of products like: $J_{ij}J_{ik}$, in which the efferent neuron i is the same in both synaptic efficacies. In general, the variables J_{ij} and J_{ik} are correlated. In Appendix A we show that the variances are

minimal when the two sets of Js are assumed independent. In that case, it is shown in the appendix that

$$R^2(\xi) = \frac{1}{N}\left(\langle J_{ij}^2 \xi_j^2\rangle_\xi - \langle h_i^p\rangle_\xi^2\right) = \frac{1}{N}\left[\sum_{\tilde{\xi}} \rho_{\tilde{\xi}} \sum_J J^2 \tilde{\xi}^2 \rho_J^p(\xi,\tilde{\xi}) - \langle h_i^p\rangle_\xi^2\right] \quad (2.4)$$

Retrieval is possible if the ratio S/R is large enough. If one requires that the probability of an error on any neuron tends to zero with increasing N, then the square of the signal-to-noise ratio must grow at least as $\log N$ (see, e.g., Weisbuch and Fogelman-Soulié 1985).

3 The Logarithmic Constraint

In the wide class of learning processes we consider below, there is always a sequence of synaptic transitions, on any given synapse, that can bring the synapse from any one of its stable states to any other state. The corresponding stochastic process is, therefore, irreducible (see, e.g., Cox and Miller 1965). In that case the matrix M has a single eigenvalue 1. Writing equation 1.1 in terms of the eigenvalues, λ_α, of M, one has

$$\rho_J^p(\xi,\tilde{\xi}) = \sum_K \rho_K^1(\xi,\tilde{\xi}) M_{KJ}^{p-1} = \rho_J^\infty + \sum_{\alpha>1} \lambda_\alpha^{p-1} \sum_K \rho_K^1(\xi,\tilde{\xi}) u_K^\alpha v_J^\alpha \quad (3.1)$$

where u^α and v^α are, respectively, the right and the left eigenvectors associated to eigenvalues λ_α. For an ergodic process, we have $\lambda_1 = 1 > \lambda_2 = \lambda_M \geq \lambda_3 \geq \cdots \lambda_n$ (Cox and Miller 1965). Note that the terms multiplying λ_α^{p-1} for $\alpha > 1$ depend on the initial conditional distribution and on the eigenvectors of M, corresponding to λ_α. They are independent of p and N.

Substituting ρ in equation 2.1 one finds

$$\langle h_i^p\rangle_\xi = h_\infty + \sum_{\alpha>1} \lambda_\alpha^{p-1} F_\alpha(\xi) \qquad (\xi = \zeta_1, \zeta_2) \quad (3.2)$$

where h_∞ is the term due to the asymptotic part of the distribution ρ_J^∞:

$$h_\infty = \langle J\xi\rangle_\infty = \sum_{\{\tilde{\xi}\}} \rho_{\tilde{\xi}} \sum_{\{J\}} J\tilde{\xi} \rho_J^\infty$$

The coefficients F_α can be read by substituting equation 3.1 in equation 2.1. They are independent of N and of p. When $\lambda_2 (= \lambda_M)$ dominates, that is,

$$\lim_{p\to\infty} \left(\frac{\lambda_\alpha}{\lambda_2}\right)^p = 0 \qquad \forall \alpha > 2$$

one can write for large p

$$\langle h_i^p\rangle_\xi = h_\infty + \lambda_M^{p-1} F_2(\xi) \qquad (\xi = \zeta_1, \zeta_2)$$

Calculating S by substituting equation 3.2 in equation 2.2, the asymptotic part h_∞ cancels, leading to

$$S = \sum_{\alpha>1} [\lambda_\alpha(\mathcal{P})]^{p-1} \cdot C_\alpha(\mathcal{P}) \quad (3.3)$$

where $C_\alpha(\mathcal{P})$ are differences of the corresponding coefficients F_α in equation 3.2, and $\mathcal{P}$ represents, schematically, their dependence on the set of parameters describing the learning dynamics. For fixed $\mathcal{P}$, λ_M dominates and

$$S = [\lambda_M(\mathcal{P})]^{p-1} \cdot C_2(\mathcal{P})$$

in which C_2 and λ_M depend on N only via an eventual dependence of one of the parameters that affect the learning dynamics (e.g., coding level of patterns, transition probabilities, presentation rate, number of stable synaptic states).

The uncorrelated part (the lower bound, see Appendix A) of the variances of the two distributions of neuronal inputs, h, are given by equation (2.4). Each of the variances has two contributions:

$$\langle J_{ij}^2 \xi_j^2\rangle_\xi = \sum_{\tilde{\xi}} \rho_{\tilde{\xi}} \sum_J J^2 \tilde{\xi}^2 \rho_J^p(\xi,\tilde{\xi}) = \sum_{\tilde{\xi}} \rho_{\tilde{\xi}} \sum_J J^2 \tilde{\xi}^2 \rho_J^\infty + \sum_{\alpha>1} \lambda_\alpha^{p-1} G_\alpha(\xi) \quad (3.4)$$

and

$$\langle h_i^p\rangle_\xi^2 = h_\infty^2 + 2h_\infty \sum_{\alpha>1} \lambda_\alpha^{p-1} F_\alpha(\xi) + \left[\sum_{\alpha>1} \lambda_\alpha^{p-1} F_\alpha(\xi)\right]^2 \quad (3.5)$$

Again, if $\lambda_2 = \lambda_M$ dominates, then

$$\begin{aligned} \langle J_{ij}^2 \xi_j^2\rangle_\xi &= \langle J^2\xi^2\rangle_\infty + \mathcal{O}(\lambda_M^{p-1}) G_2(\xi) \\ \langle h_i^p\rangle_\xi^2 &= h_\infty^2 + 2\lambda_M^{p-1} F_2(\xi) h_\infty \end{aligned}$$

The dependence on ξ is contained in the functions G_2, F_2. For $p \to \infty$ all the terms that are multiplied by λ_M^{p-1} disappear and only the asymptotic part survives. So the signal-to-noise ratio behaves as

$$\frac{S^2}{R^2} = \lambda_M^{2(p-1)} N \cdot C(\mathcal{P}) \quad (3.6)$$

in which

$$C(\mathcal{P}) = \frac{[C_2(\mathcal{P})]^2}{\langle J^2\xi^2\rangle_\infty - h_\infty^2}$$

If we impose that in the limit $N \to \infty$ the ratio S^2/R^2 grows at least as $\log N$, then we obtain a bound on p:

$$p_c < \frac{1}{-2\log\lambda_M(\mathcal{P})} \log\left[\frac{N \cdot C(\mathcal{P})}{\log N}\right] \quad (3.7)$$

This result makes sense, of course, only if the argument of the logarithm is greater than unity. Or that

$$NC(\mathcal{P}) > \log N \tag{3.8}$$

Setting p=1 in equation 3.6, this condition is seen to be equivalent to the condition that the ratio of signal to noise will allow the recall of the most recently learned pattern (λ_M is strictly less than 1).

In fact, the result 3.8 is a gross overestimate. The correlations mentioned above and discussed in Appendix A can make the variances remain finite as N becomes large. The increase in p_c with N is all due to the fact that the noise decreases as N^{-1}. Moreover, when the noise does not decrease with increasing N, the product $NC(\mathcal{P})$ does not increase with N. Hence the condition 3.8 can never be satisfied. As we proceed to show in what follows, the escapes from the tight storage constraint on p_c are effective also when the correlations are included.

4 Possible Escapes

The logarithmic constraint on the number of retrievable patterns concerns a very wide class of networks with dynamic synapses. However, the form of p_c, equation 3.7, suggests possible escapes. If one allows the parameters, $\mathcal{P}$, contained in λ_M to vary with the size of the network, so that $\lambda_M \to 1$, then it is possible to go beyond the logarithmic constraint. The corresponding variation of $C(\mathcal{P})$, limits the space of variation of the parameters $\mathcal{P}$. Specifically, if λ_M has the form

$$\lambda_M = 1 - x(\mathcal{P}) \tag{4.1}$$

and the dependence of $\mathcal{P}$ on N makes $x \to 0$ in the limit of large N, while the constraint 3.8 is respected, then

$$p_c \sim x^{-1}. \tag{4.2}$$

As mentioned at the end of the last section, if the constraint is not satisfied, there is no way to improve memory. Making λ_M tend toward unity can, at best, prolong the trace of the first imprinted pattern. But when the constraint is violated, there is no trace to maintain. Fortunately, in the memory optimizing cases to be considered the correlations contribute a negligible amount to the variances (see, e.g., Appendix A).

We have considered four types of parameters $\mathcal{P}$ that affect the learning dynamics and that may depend on N:

- **Speed of pattern presentation.** If the number of stimuli presented to the network in the interval of a single transition between the synaptic states increases to m, the storage capacity is multiplied by m (Amit and Fusi 1992). Imposing a minimal rate of presentation seems rather artificial so we shall look for those types of remedies which improve the worst case: low rate presentation ($m = 1$).

- **Stochastic refresh mechanism.** The transition probabilities of a synapse for given input can be made to decrease with N.
- **Coding level of the patterns.** The fraction of information carrying bits per stimulus can be made to decrease with increasing N.
- **Number of synaptic states.** The number of stable states per synapse, n, can increase with N.

But when $p \sim \mathcal{O}(x^{-1})$ we have $(\lambda_M)^p \to Const \neq 0$ as $x \to 0$. In that case one must reexamine the dominance of $\lambda_M = \lambda_2$ in the expansion equation 3.1. In fact, usually a whole set of eigenvalues $\lambda_\alpha \to 1$ and $(\lambda_\alpha)^p \to K_\alpha \neq 0$ in this limit (see, e.g., Section 7). The part of $\rho_f^p(\xi, \tilde{\xi})$ corresponding to λ_1 remains distinguished from the contributions due to the other $\lambda_\alpha \to 1$, because it is the only part that is independent of the first learned pattern.

The appearance of several eigenvalues for which $(\lambda_\alpha)^p \to K_\alpha \neq 0$ implies that sums over eigenvalues, such as in equations 3.3, 3.4, and 3.5 separate into two parts: one part running over all the eigenvalues which tend to 1 and a part that includes all the lower eigenvalues and hence tends to zero. Since we have taken $p \sim x^{-1}$, the remaining sum may depend on x and effectively change the factor $C(\mathcal{P})$ in equation 3.7 or 3.8, thus possibly modifying the constraint on the range of variation of the learning parameters $\mathcal{P}$. In at least one such case, the case studied in Section 7, we find that no such change is induced by the degeneracy of the eigenvalues in the limit.

5 Stochastic Learning of Sparsely Coded Patterns

The most interesting results appear in the case of 0–1 neurons, with a low mean fraction f of 1s and synapses with 2 stable states (J_-, J_+). An imposed stimulus can produce the following transitions at a synapse:

- If $J_{ij} = J_-$ and the new stimulus activates the associated pair of neurons (i.e., ξ_i^μ=ξ_j^μ=1), then a transition $J_- \to J_+$ occurs with probability q_+. So upon each presentation of a new pattern the probability of potentiation is $f^2 q_+$.
- If $J_{ij} = J_+$ and the stimulus contains a mismatched pair of activities, then the transition probabilities for a depression $J_+ \to J_-$ are $q_-(10)$ for $\xi_i^\mu = 1, \xi_j^\mu = 0$ and $q_-(01)$ when $\xi_i^\mu = 0, \xi_j^\mu = 1$. The transition probabilities $q_-(10)$, $q_-(01)$ may differ. Denoting $q_- = q_-(10) + q_-(01)$ we have that the total probability of a synaptic depression is $f(1-f)q_-$.
- A pair of inactive neurons leaves the corresponding synapse unchanged.

The resulting transition matrix is

$$M = \begin{bmatrix} 1-f(1-f)q_- & f(1-f)q_- \\ f^2q_+ & 1-f^2q_+ \end{bmatrix} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} \tag{5.1}$$

where $a = f(1-f)q_-$, $b = f^2q_+$. The two eigenvalues are 1 and:

$$\lambda_M = 1 - f^2q_+ - f(1-f)q_- \tag{5.2}$$

The asymptotic distribution, the left eigenvector belonging to the eigenvalue 1, is

$$\rho^\infty = \left(\frac{b}{a+b}, \frac{a}{a+b}\right) = (p_+, p_-) \tag{5.3}$$

where $p_+ = b/(a+b)$ is the fraction of synapses with value J_+. Note that the Willshaw (1969) model has $J_+ = 1, J_- = 0$ q_-=0. Hence, $a = 0$ and consequently $\rho^\infty = (1,0)$: all synapses become 1.

For the present case, since $\rho^p_{J_-} = 1 - \rho^p_{J_+}$, equation 2.3 becomes

$$\langle h\rangle_{+1} = (J_+ - J_-)f\rho^p_{J_+}(1,1) + J_-f \tag{5.4}$$

$$\langle h\rangle_0 = (J_+ - J_-)f\rho^p_{J_+}(0,1) + J_-f \tag{5.5}$$

The conditional probabilities are given by equation 3.1 as

$$\rho^p_{J_+}(1,1) = \lambda_M^{p-1}\sum_K \rho^1_K(1,1)u_K v_{J_+} + \rho^\infty_{J_+} = \lambda_M^{p-1}\left[\rho^1_{J_+}(1,1) - p_+\right] + \rho^\infty_{J_+}$$

$$\rho^p_{J_+}(0,1) = \lambda_M^{p-1}\sum_K \rho^1_K(0,1)u_K v_{J_+} + \rho^\infty_{J_+} = \lambda_M^{p-1}\left[\rho^1_{J_+}(0,1) - p_+\right] + \rho^\infty_{J_+}$$

where the eigenvectors corresponding to λ_M are given by $u_K = (p_-, -p_+)$ and $v_K = (1,-1)$.

Assuming that one starts from asymptotic distribution, the conditional probabilities following the presentation of the oldest pattern ξ^1 are

$$\rho^1_{J_+}(1,1) = p_+ + p_-q_+\,, \qquad \rho^1_{J_+}(0,1) = p_+(1-q_-) \tag{5.6}$$

When calculating the signal, the asymptotic parts cancel and the leading term, is proportional to λ_M^{p-1}.

$$S = \lambda_M^{p-1}(J_+ - J_-)(q_+p_- + q_-p_+)f \tag{5.7}$$

Note again that for the Willshaw model $\rho^1 = \rho^\infty$ and S=0, that is, no learning is possible on top of the asymptotic distribution.

For the calculation of the uncorrelated part of the noise, equation 2.4, we need $\langle h^2\rangle$, which is the same as $\langle h\rangle$ with J_+ replaced by J_+^2 and J_- by J_-^2. One finds that

$$R^2 = \frac{f}{N}\left[p_+J_+^2 + p_-J_-^2 - (p_+J_+ + p_-J_-)^2 f + \mathcal{O}(\lambda_M^{p-1})\right] \tag{5.8}$$

For small f we keep only terms of leading order in f and, for large p, the signal-to-noise ratio is

$$\frac{S^2}{R^2} = \lambda_M^{2(p-1)}Nf\left(1 - \frac{J_-}{J_+}\right)^2 \frac{(q_+p_- + q_-p_+)^2}{p_+ + p_-\left(\frac{J_-}{J_+}\right)^2} \tag{5.9}$$

5.1 Extremal Cases and the Return of Optimal Storage.

5.1.1 Lowest Coding Level. First we take the coding level $f \sim \log N/N$ (as in Willshaw 1969) keeping the transition probabilities q_+ and q_- fixed and both different from zero. From equation 5.2 we read that $\lambda_M \sim 1 - x$ (equation 4.1) with

$$x = f^2q_+ + f(1-f)q_- = \mathcal{O}(f) = \mathcal{O}(\log N/N)$$

and, according to equation 4.2,

$$p_c = \mathcal{O}\left(\frac{N}{\log N}\right)$$

In fact, $f \sim \log N/N$ is as low as f is allowed to become without violating the bound (3.8). Moreover, even the above result for p_c is too high. The reason is that when q_+ and q_- are fixed, the correlation term, Appendix A, overpowers the leading uncorrelated part when p_c goes above $N/(\log N)^2$. In other words, this network performs much worse than Willshaw's (1969), which for the same coding level gives $p_c \sim N^2/\log^2 N$. This is a price for continual learning.

5.1.2 Optimal Storage Recovered. The optimal performance can be recuperated if we take $f \sim \log N/N$ and the transition probability $q_- = fq_+$. Provided the bound (3.8) is not violated, according to equation 4.2, since now x of equation 4.1, is $\sim (\log N/N)^2$, one has the optimal storage

$$p_c \sim \left(\frac{N}{\log N}\right)^2$$

if q_+ does not tend to zero.

To verify that the retrieval bound, (3.8), is respected one first notes that in this case the part of the noise due to correlations is negligible. It is of magnitude pf^3 relative to the uncorrelated part (see, e.g., Appendix A). It is therefore sufficient to read $C(\mathcal{P})$ from equation 5.9 and to substitute it in equation 3.8. In the present case the asymptotic fractions p_+ and p_- of J_+ and J_-, respectively, are finite. The only strong N dependence in $C(\mathcal{P})$ is in f and hence the constraint reduces to $Nf = \mathcal{O}(\log N)$.

5.1.3 Intermediate Cases. One could attempt to trade off some of the N dependence of f for an N dependence of q_+, which has been assumed finite in the limit of large N. The constraint on $C(\mathcal{P})$ implies that if

$$f = \left(\frac{\log N}{N}\right)^\beta$$

then

$$q_+^2 = \left(\frac{\log N}{N}\right)^{1-\beta}$$

with $\beta \in [0,1]$ ($f < 1$ implies that $\beta > 0$ and $q_+ < 1$ gives the upper bound $\beta < 1$). For x of equation 4.1 we have

$$x \sim f^2 q_+ \sim \left(\frac{\log N}{N}\right)^{\frac{1}{2}+\frac{3}{2}\beta}$$

and hence

$$p_c < \frac{1}{x} = \left(\frac{N}{\log N}\right)^{\frac{1}{2}+\frac{3}{2}\beta}$$

The discussion in Appendix A shows that in the part of the intermediate regime in which $\beta > \frac{1}{3}$, the correlated part of the noise is negligible. The case β=0, that is, fixed finite coding level f, reproduces the result of Tsodyks (1990), with a capacity of $\mathcal{O}(\sqrt{N})$.

6 Simulations

We have carried out extensive simulations to test the predictions of the theoretical estimates in the most extreme case, that of optimal storage in 2-state synapses and 0–1 neurons. In fact, to make the test of the theory more stringent, we have tested separately the asymptotic behavior of the signal and the noise. In the simulations the parameters were set as follows:

$$J_+ = 1,\ J_- = 0, \quad f(N) = A\frac{\log N}{N}, \quad q_+ = 1, \quad q_- = f \tag{6.1}$$

with fixed $A = 4$. The signal and the noise are estimated for each choice of parameters N and p in the following way:

A sequence of $N_p = 500 + p$ random N-bit words is generated, the stimuli to be learned. p is the maximal age of a pattern to be tested. Each word is generated by assigning 1's, at random, with probability f. All $500+p$ patterns are presented consecutively to the network. Upon the arrival of each pattern ξ^μ, learning takes place, modifying the synaptic matrix according to the learning rule described at the beginning of Section 5. Following the learning of ξ^μ ($p < \mu < p+500$) the state of the network is set to the pattern of age p, $s_i = \xi_i^{\mu-p}$, that is, the stimulus to be retrieved. Then, with the new synaptic matrix J_{ij}^μ, we calculate the average of the postsynaptic input over the foreground and the background neurons in order to estimate the conditional mean of equation 2.3, that is,

$$\langle h^p \rangle_\zeta(\mu) = \frac{1}{N_\zeta} \sum_{i:(s_i=\zeta)} \sum_{j \neq i} J_{ij}^\mu \xi_j^{\mu-p}$$

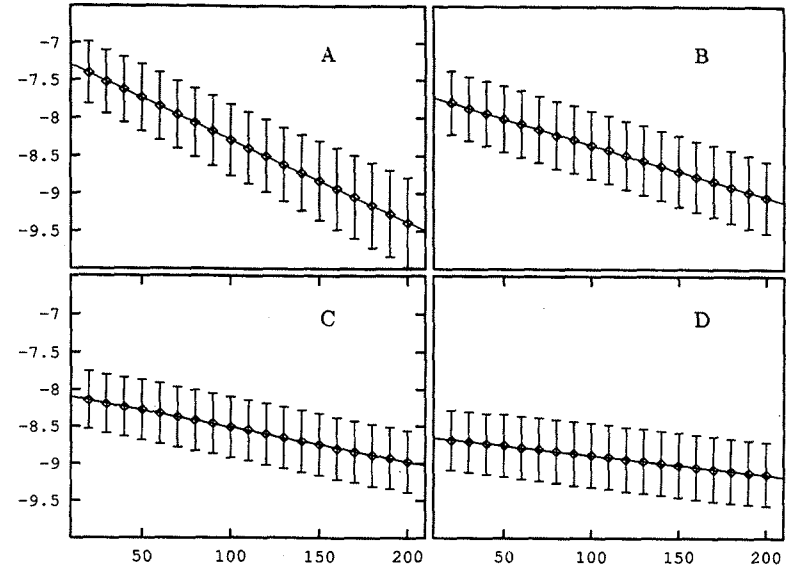

Figure 1: Logarithm of signal vs. number of memories p for fixed N: (A) $N = 600$, (B) $N = 800$, (C) $N = 1000$, (D) $N = 1400$. The lines are a linear fit of the mean signals. The slopes $a_1(N)$ are reported in Table 1. Error bars are rms in measured signals.

where the index i runs over all neurons for which $\xi_i^{\mu-p} = \zeta(=0,1)$; N_ζ is the number of neurons with $s_i = \zeta$(=0,1) in the pattern presented. From these data we compute the square of the signal as the average over all presentations, that is,

$$S^2 = \frac{1}{N_p} \sum_{\mu=1}^{N_p} [\langle h^p \rangle_1(\mu) - \langle h^p \rangle_0(\mu)]^2$$

And the noise R^2 is calculated as half the sum of the standard deviations of h around $\langle h^p \rangle_1$ and $\langle h^p \rangle_0$.

These results are then compared to the theoretical estimates. In particular we have tested the dependence of S^2 on N for fixed p and its dependence on p for fixed N. The theoretical expectations for the square of the signal, equation 5.7, are

$$S^2(p,N) \simeq [\lambda_M(N)]^{2(p-1)} f^2(N) \left[\frac{2-f(N)}{2-3f(N)}\right]^2 \tag{6.2}$$

where $f(N)$ is defined in equation 6.1. With the present choice of parameters

$$\lambda_M = 1 - 3f^2(N) + 2f^3(N)$$

The theoretical upper bound estimate for the noise can be obtained from equation 5.8. One has

$$R^2(p,N) = \frac{1}{N}\left\{f(N)p_+(N) + \lambda_M^{p-1}\frac{f(N)}{2}[1-p_+(N)] - \mathcal{O}(f^2)\right\}$$
$$< \frac{f(N)}{2N}[1+p_+(N)] \tag{6.3}$$

If equations 6.2 and 6.3 are verified in the regime of the asymptotic behavior in N, then the number of storable and retrievable patterns can grow as $N^2/\log^2 N$. Indeed, as long as p is bounded by this value, there exists a threshold that separates the depolarization of neurons that should be active from those that should be quiescent.

Equation 6.2 is written in the form

$$y_1 = \log S^2 = a_1(N)p + b_1(N) \tag{6.4}$$

with

$$a_1(N) = 2\log[1 - 3f^2(N) + 2f^3(N)] \tag{6.5}$$

The four insets in Figure 1 present $\log S^2$ vs p for $N = 600$, 800, 1000, and 1400. The straight lines are a fit of the mean signals by equation 6.4. From these fits we find values for $a_1(N)$ that are compared in Table 1 to the theoretical values given by equation 6.5 for several values of N. The agreement represented in the table implies that in the entire range of values of N and of p tested in the simulations one is already in the asymptotic regime for the behavior in N and p. Hence the fact that in this region $S^2/R^2 > \log N$ implies, in turn, storage capacity quadratic in N.

The behavior of S^2 vs N, for the same value of A, is presented in Figure 2 where S^2 is plotted as function of N for four different values of p (20, 30, 40, 60). The continuous line represents the theoretical estimate while the points are simulation results. The agreement improves with increasing N although, even for small N, the theoretical lines pass through the errorbars. It is worth noting that in case D the nonmonotonic behavior around $N = 400$ is captured by the theory.

Finally the upper bound on R^2 is tested in Figure 3. In particular, R^2 is plotted as a function of p for $N = 600$, 800, 1000, and 1400. The noise tends to its asymptotic value, and is always below the straight dotted line which represents the upper bound (equation 6.3). The value of the upper bound decreases with increasing N and R^2 approaches its asymptotic limit more slowly when N is larger. This is due to the fact that for large N λ_M is closer to 1, and the correction to asymptotic distribution goes to zero more slowly (see equation 6.3).

Table 1: Testing the Asymptotic Regime.[a]

N	theoretical a_1	a_1 from simulations
400	−0.0208	−0.0218 ± 0.0029
500	−0.0144	−0.0148 ± 0.0022
600	−0.0106	−0.0110 ± 0.0021
700	−0.0082	−0.0086 ± 0.0019
800	−0.0066	−0.0068 ± 0.0019
900	−0.0054	−0.0056 ± 0.0019
1000	−0.0045	−0.0045 ± 0.0017
1100	−0.0038	−0.0040 ± 0.0017
1200	−0.0033	−0.0033 ± 0.0016

[a]Comparison between theoretical $a_1(N)$, equation 6.4, and the value measured in simulations.

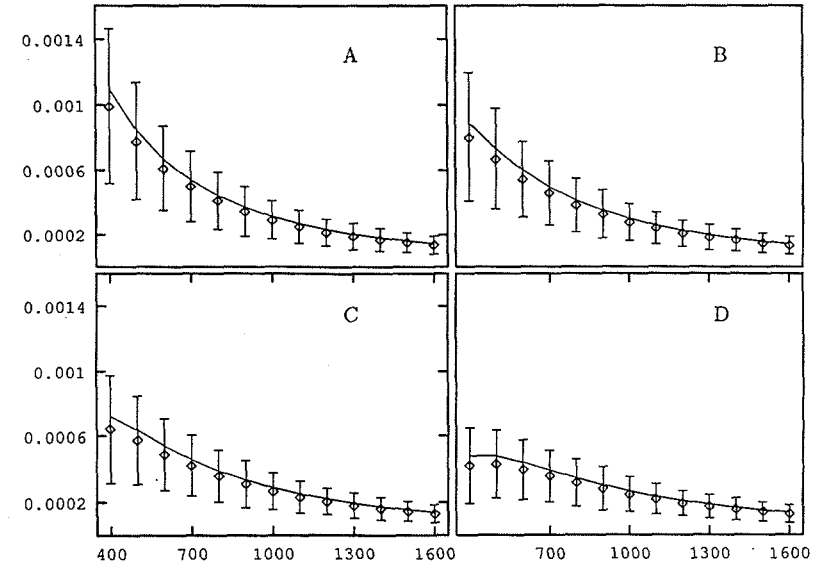

Figure 2: S^2 vs N for several values of p: (A) $p = 20$, (B) $p = 30$, (C) $p = 40$, (D) $p = 60$. Dots are simulations results. The continuous line is the theoretical prediction (equation 6.2). Note the improvement of agreement with increasing N.

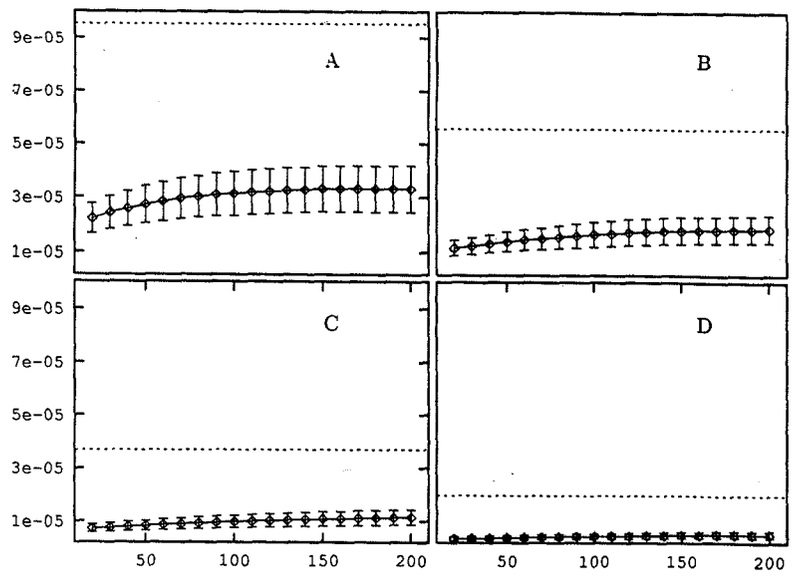

Figure 3: Simulation results for R^2 vs. p: (A) $N = 600$, (B) $N = 800$, (C) $N = 1000$, (D) $N = 1400$. The horizontal lines are the theoretical upper bound (equation 6.3). When p grows then R^2 approaches exponentially its asymptotic value.

7 Multistate Synapses and ±1 Neurons

The second example we consider, to demonstrate the dependence of the performance of an autoassociative network on the number of stable synaptic states, is a network of ±1 neurons ($\zeta_1 = 1$, $\zeta_2 = -1$) and synapses with n stable states:

$$J_m = -1 + \frac{2m}{n-1}, \qquad m = 0, \ldots, n-1 \tag{7.1}$$

Each pattern ξ_i^μ is a random word of N ±1 bits chosen independently and with equal probability [$\Pr(\xi = 1) = \Pr(\xi = -1) = 1/2$]. Upon presentation of a pattern a synapse is potentiated ($J_m \to J_{m+1}$) with probability q if the source $\xi_i^\mu \xi_j^\mu$=1 or depressed with the same probability ($J_m \to J_{m-1}$) if $\xi_i^\mu \xi_j^\mu = -1$. If a synapse is at one of its extreme limits and is pushed on, its value is unchanged. So in the process of the presentation of patterns a synapse undergoes a random walk between two reflecting barriers. Note that in this model a stimulus communicates information to all N neurons and qN^2 synapses are modified by each stimulus.

The stochastic transition matrix M_{KJ} is tridiagonal with $q/2$ along the two side diagonals: $1-q$ along the main diagonal, except in the first and last positions where it is $1 - q/2$. Since this matrix is symmetric, its right and left eigenvectors are identical and hence the asymptotic distribution is uniform, that is

$$\rho_J^\infty = \frac{1}{n} \qquad \forall J$$

and hence $h_\infty = 0$ due to the ±1 symmetry.

The full set of eigenvalues of M is

$$\lambda_\alpha = 1 - 2q \sin^2 \frac{\pi\alpha}{2n} \tag{7.2}$$

with $\alpha = 0, \ldots, n-1$ ($\lambda_0 = 1$). Its second largest eigenvalue $\lambda_1 = \lambda_M$ is

$$\lambda_M = 1 - 2q \sin^2 \frac{\pi}{2n}$$

and the corresponding eigenvector is

$$v_k = c \cos \frac{\pi k}{n-1}, \qquad k = 0, \ldots, n-1 \tag{7.3}$$

where, for large n, the constant c behaves as $\sqrt{2/n}$. The contribution $\langle h \rangle_{+1}$ to the signal is

$$\langle h \rangle_{+1} = \sum_J J \left[\rho_J^p(1,1) - \rho_J^p(1,-1) \right]$$

where the index J runs over all the stable synaptic values $J = J_m, m = 0, \ldots, n-1$ (equation 7.1). Recall that $\langle h \rangle_{-1} = -\langle h \rangle_{+1}$. Substituting the expression for the conditional distributions one finds

$$\langle h \rangle_{+1} = \lambda_M^{p-1} \sum_J J \sum_K v_K v_J \left[\rho_K^1(1,1) - \rho_K^1(1,-1) \right] + \mathcal{O}(\lambda_3^{p-1}) \tag{7.4}$$

The difference of the two conditional distributions, following the presentation of the oldest pattern ξ^1, is

$$\rho_K^1(1,1) - \rho_K^1(1,-1) = \left(\frac{1}{n} 2q, 0, \ldots, 0, -\frac{1}{n} 2q \right) \tag{7.5}$$

This form follows from the observation that when $\xi_i^1 \xi_j^1 = 1$ is presented to the uniform asymptotic distribution it leaves the probability of all states invariant except the two extreme ones: The lowest state that loses a fraction q of its occupation, becoming $(1-q)/n$ and the top one which gains the balance and becomes $(1+q)/n$. $\xi_i^1 \xi_j^1 = -1$ does the opposite, producing the following two conditional distributions:

$$\begin{aligned} \rho_K^1(1,1) &= \frac{1}{n}(1-q, 1, \ldots, 1, 1+q) \quad \rho_K^1(1,-1) \\ &= \frac{1}{n}(1+q, 1, \ldots, 1, 1-q) \end{aligned} \tag{7.6}$$

Hence, taking $2\langle h\rangle_{+1}$ of equation 7.4 and substituting equations 7.5 and 7.3 the signal becomes

$$S \simeq \frac{q}{n}\lambda_M^{p-1}C \tag{7.7}$$

where the constant C is independent of q, n, N, p. Note that the signal decreases with increasing n. This is a consequence of the fact that the process has a uniform asymptotic distribution of values. Even after the presentation of a single pattern, on the background of the asymptotic distribution, the signal will decrease as $1/n$.

The noise can be calculated using equations 3.4 and 3.5:

$$R^2(\xi) = \frac{1}{N}\left[\sum_{m=0}^{n-1} J_m^2 \rho_{J_m}^{\infty} + \mathcal{O}(\lambda_M^{p-1})\right]$$

In this case, due to the $\pm$ symmetry of the process, the contributions of the synaptic correlations to the noise vanish identically. The dependence on ξ is contained in the term $\mathcal{O}(\lambda_M^{p-1})$, which vanishes as $p \to \infty$. Hence

$$R^2 = \frac{1}{N}\left[\sum_{m=0}^{n-1} J_m^2 \rho_{J_m}^{\infty}\right]$$

where J_m is given by equation 7.1. Substituting J_m one has

$$R^2 \sim \frac{8}{nN}\sum_{m=0}^{n-1}\left(-1 + \frac{2m}{n-1}\right)^2 \tag{7.8}$$

which does not depend on n, because the sum grows linearly in n.

Hence the final signal-to-noise ratio behaves like

$$\frac{S^2}{R^2} = \frac{q^2}{n^2} N \lambda_M^{2(p-1)} K \tag{7.9}$$

where K is the ratio of C^2 to the part of R^2 that does not depend on N. Hence

$$p_c < \frac{1}{-2\log\lambda_M}\log\left(\frac{q^2 N}{n^2 \log N}\right) \tag{7.10}$$

Again, if q and n are fixed, the capacity is logarithmic.

On the other hand, if q and n are chosen so that $q/n^2 \to 0$ then λ_M tends to 1 and x of equation 4.1 is

$$x = \frac{\pi^2 q}{2n^2}$$

Equation 4.2 gives

$$p_c < \frac{n^2}{\pi^2 q}\log\left(\frac{q^2 N}{n^2 \log N}\right) \tag{7.11}$$

7.1 Extremal Case. The constraint on the allowed variation of n and q, equation 3.8, is equivalent to the requirement that the log in equation 7.11 be positive. Writing the probability q in the form

$$q^2 = \left(\frac{\log N}{N}\right)^{\beta}$$

We must have $\beta \geq 0$, since q must lie in the interval $[0, 1]$. Moreover, the constraint implies that

$$n^2 = \left(\frac{N}{\log N}\right)^{1-\beta}$$

and since $n > 1$, $\beta < 1$. Substituting n and q in equation 7.11 we find

$$p_c < \left(\frac{N}{\log N}\right)^{\frac{1}{2}+\frac{1}{2}\beta} \tag{7.12}$$

Since β is restricted to the interval [0,1], the number of stable synaptic states cannot surpass $\sqrt{N}$. If n becomes larger, the network is no longer a palimpsest and all memory is destroyed together. When n reaches this limit ($\beta = 0$) the number of retrievable memories is proportional to N (see, e.g., Parisi 1986). The price is that the number of synaptic states is not a property of the synapse, it increases with the size of the network.

If one introduces a stochastic transition mechanism with $q \neq 1$ ($\beta > 0$) then it is possible to store more than $\sqrt{N}$ patterns. When β varies from 1 to 0 the process interpolates between $p \simeq \sqrt{N}$ to $p \simeq N$.

7.2 The Role of the Other Eigenvalues. As was discussed at the end of Section 4, when $\lambda_M \to 1$ the contribution of the other eigenvalues must be reexamined. Equation 7.2 for the eigenvalues implies that all n of them tend to 1 as $n \to \infty$ or $q \to 0$. Nevertheless, we show in Appendix B that the dependence of both the signal and the noise on n and on q remains unchanged.

8 Discussion

We have tried to open a discussion of the consequences of synaptic dynamics that may be taking place in a network that receives a temporally unconstrained stream of stimuli and maintains the same neural and synaptic dynamics whether the network is engaged in computation or in learning new memories. The requirement that memory be preserved in

the synapses for long times has induced us to postulate that synapses have finite sets of states that are stable. In between such states the synapse is assumed to be able to vary continuously, but the analog values can be maintained only for short times and their main role is to allow a synapse, based on the neural activity in the neurons connected by it, to cross thresholds for transitions between neighboring stable states.

Whether biological synapses have such ladders of stable states is a question of neurophysiology and biochemistry. Given recent progress in measuring synaptic efficacies between single pyramidal neurons (Mason *et al.* 1991), the neurophysiological test may soon be feasible. On the other hand, in electronic implementations of unsupervised neural networks this mechanism has proved very natural and effective. The stable states of a synapse are essentially (see, e.g., Amit *et al.* 1992; Badoni *et al.* 1992) an asynchronous refresh mechanism operating on some capacity associated with the synapse. The simplicity of implementation, the economy in means, and the accessibility to analysis should make this scheme rather attractive.

In studying the retrievability of learned patterns only the existence of a potential threshold has been considered. We have ignored the possibility that for different stimuli this threshold may differ. In fact, it does, mostly due to fluctuations in the number of active neurons from pattern to pattern. We have noticed elsewhere (Amit and Brunel 1993) that this problem is naturally overcome by an unstructured inhibition reacting in proportion to the total excitatory activity.

Another issue mentioned but not developed concerns the possible analog nature of the information coding in the stimuli. It has been raised in Section 1 in connection with the origin of the nondeterministic nature of the synaptic transition given the same pair of digitally coded information bits. In the discussion of the retrieval we have considered only the digital representation of the stimuli that had been learned. If in fact the fluctuating nature of the transition probabilities is related to the fluctuations of neuronal activity variables, such as spike rates, one must test also the retrieval of patterns that have fluctuating variables. We have not done this, either theoretically or in the simulations, but we believe that the modification should be minor. The reason is that for the digital coding to make sense it must represent approximately the analog variables. In other words, a neuron with a 0 digital code will have low frequency and one with a digital code of 1 will have high frequency. Thus the difference between the presentation of the analog vs. the digital pattern for retrieval can be considered as noise on the incoming stimulus to be retrieved.

We have emphasized the issue of the palimpsest behavior of the networks. In the present context this type of behavior is quite natural. One may wonder whether experience indicates that brain functions as a palimpsest. We are not familiar with any direct evidence that this is the case, yet the issue is not moot. First, if the storage capacity of any

cortical module is of relevance, clearly the behavior of that module near capacity becomes important. At that point it is pertinent to raise the question of whether it behaves as a palimpsest or not. Experience does not produce the impression that old memories are replaced by new ones. In fact, often one has the opposite impression, that is, that old memories never die. Yet it should be kept in mind that the theoretical treatment presented here has dealt with strings of stimuli that are uncorrelated. The repetition of some subclasses of stimuli in the process of learning may create privileged memories. How this is included in a theoretical framework we leave as an open question.

What seems important to realize in this context is that it is quite possible that a module will receive a very rich stream of stimuli. Since there is no dynamic distinction between those that should be learned and those that are transient, all make some modification of the synaptic structure. It may be the case that most of what enters the module is noise and hence that what is learned *is* learned on the background of an asymptotic distribution of synaptic values. This is the deeper sense of palimpsest behavior in our context.

This connects with another question: what is the dynamics for learning correlations in the input stream? Such correlation may be of two types: there may be correlations in the spatial activity distribution of patterns in the afferent sequence. Or it may be that the system manages to learn temporal correlations in the sequence, as is implied by the experiments of Miyashita (1988; Griniasty *et al.* 1992). In both cases there is a need for an extension of the techniques presented here. It appears that in some cases such extensions are not unsurmountable (Brunel 1993).

Finally, one may be struck by the discrepancy between the tight storage bound that we find for networks with fixed parameters and the results on the capacity of networks with ± 1 synapses of Amaldi and Nicolis (1989), Gutfreund and Stein (1990), and Krauth and Mezard (1989), which give a capacity linear in the number of neurons. Our conclusion is that there is no local learning algorithm that can lead to those matrices. Nonlocality is invoked in a double sense; it is spatial as well as temporal—spatial, because one needs more than the two activities imposed by the stimulus on the pair of neurons connected by the synapse to be modified and temporal, because one must know all the stimuli simultaneously while deciding if a modification is acceptable or not.

Acknowledgments

We are indebted to Profs. Giorgio Parisi and Fabio Martinelli for advice concerning random walks between reflecting barriers and to Nicolas Brunel for discussions. We have a special debt to Prof. Yali Amit for pointing out to us the role of synaptic correlations that we overlooked in a previous version of this article.

Appendix A

The full expression for the variance of the neuronal input about its mean in one of the classes is

$$\begin{aligned} R_\xi^2 &= \left\langle (h_i - \langle h_i\rangle_\xi)^2 \right\rangle_\xi = \left\langle \frac{1}{N^2}\sum_{j\neq i}\sum_{k\neq i} J_{ij}J_{ik}\xi_j^1\xi_k^1 \right\rangle_\xi - \left\langle \frac{1}{N}\sum_{j\neq i} J_{ij}\xi_j^1 \right\rangle_\xi^2 \\ &= \left\langle \frac{1}{N^2}\sum_{j,k\neq i; j\neq k} J_{ij}J_{ik}\xi_j^1\xi_k^1 \right\rangle_\xi + \left\langle \frac{1}{N^2}\sum_{j\neq i}(J_{ij})^2(\xi_j^1)^2 \right\rangle_\xi - \left\langle \frac{1}{N}\sum_{j\neq i} J_{ij}\xi_j^1 \right\rangle_\xi^2 \\ &= \frac{N^2-3N+2}{N^2}\left\langle J_{ij}J_{ik}\xi_j^1\xi_k^1\right\rangle_\xi + \frac{N-1}{N^2}\left\langle J^2(\xi^1)^2\right\rangle_\xi - \frac{(N-1)^2}{N^2}\left\langle J\xi^1\right\rangle_\xi^2 \\ &= \frac{N-1}{N^2}\left[\langle J^2(\xi^1)^2\rangle_\xi - \langle J\xi^1\rangle_\xi^2\right] \\ &\quad + \frac{(N-2)(N-1)}{N^2}\left[\langle J_{ij}J_{ik}\xi_j^1\xi_k^1\rangle_\xi - \langle J\xi^1\rangle_\xi^2\right] \end{aligned} \tag{A.1}$$

The first term on the last line is the term that ignores correlations between the distributions of J_{ij} and J_{ik}. The second one, which is of order 1 as $N \to \infty$, is due to the correlations.

At this stage we can conclude that the uncorrelated contribution to the variance, the first term in equation A.1, gives a lower bound on the total. The reason is that since in the large N limit the second term dominates the variance, it must necessarily be positive. Otherwise the total variance may become negative. Hence the second term can only increase the total variance.

To calculate the first term in the second square brackets requires the conditional probability distribution $\rho(J_{ij}J_{ik} \mid \xi\tilde{\xi}_1\tilde{\xi}_2)$, where ξ is the value of ξ_i^1 common to both synapses. Thus we need the difference

$$\delta\rho = \rho(J_{ij}J_{ik}|\xi\tilde{\xi}_1\tilde{\xi}_2) - \rho(J_{ij}|\xi\tilde{\xi}_1)\rho(J_{ij}|\xi\tilde{\xi}_2) \tag{A.2}$$

To obtain the first term one has to use an equation of the type of equation 1.1, for the conditional probability that a pair of synapses with a common neuron has a given pair of values, conditioned on the values of the three neurons—i (the common one), j and k, in upon the presentation of pattern number 1. The distribution $\rho(J_1J_2 \mid \xi\tilde{\xi}_1\xi_2)$, for fixed ξ, has four values, since each of the two Js can have two values. Hence, the transition matrix, corresponding to M in equation 1.1, is a 4×4 matrix.

We have computed this matrix as well as the difference with the matrix driving the two synaptic values in the uncorrelated case for the model described in Section 5. The latter matrix is simply the outer product of the two 2×2 matrices of equation 5.1. We skip the details, which are straightforward but tedious, and summarize the results: The difference between the correlated transition matrix and the uncorrelated one is again a 4×4 matrix. For small values of f, q_-, and q_+ its elements are all dominated by the largest of the terms:

$$fq_-^2, \qquad f^2q_-q_+, \qquad f^3q_+^2 \tag{A.3}$$

When the transition matrix is raised to the power p, the contribution to the difference $\delta\rho^p$ can be estimated by

$$pM^{p-1}\delta M$$

where δM is the difference of the transition matrices in the correlated and the uncorrelated cases and M^{p-1} is the uncorrelated transition matrix iterated $p-1$ times. Both terms in the correlated contribution to the variance are proportional to f^2, since in the averages they contain two independent sums over variables ξ. Hence the estimate of the correlated contribution is f^2 multiplied by the largest of the three terms in A.3. On the other hand, the uncorrelated part of the variance is dominated by f/N for small f and large N (see, e.g., equation 5.8).

Example 1. if q_- and q_+ are fixed, as N becomes large, the leading term in the correlated part of the variance comes from the term linear in f and the uncorrelated part will dominate as long as

$$\frac{f}{N} \gg pf^3$$

Hence, in particular, when $f \sim \log N/N$ and $p = f^{-1}$, the correlated term takes over, leading to a violation of the retrieval criterion.

Example 2. The intermediate cases discussed in Section 5. Taking $q_- = fq_+$ all three terms in A.3 become of the same order: $f^3q_+^2$. Multiplying by pf^2 and comparing to f/N gives for the leading terms in the variance

$$\frac{f}{N}(1 + pNf^4q_+^2)$$

where the second term in the parentheses is due to the correlations. With the notation of Section 5, that is,

$$f = \left(\frac{\log N}{N}\right)^\beta; \qquad q_+^2 = \left(\frac{\log N}{N}\right)^{1-\beta}; \qquad p_c = \left(\frac{N}{\log N}\right)^{1/2+3/2\beta}$$

the correlation term becomes

$$pNf^4q_+^2 = N\left(\frac{\log N}{N}\right)^{\frac{1}{2}+\frac{3}{2}\beta}$$

For this term to become negligible compared to 1, we must have $\beta > 1/3$.

Appendix B

Here we verify that the result (7.12) remains unchanged when one includes the contribution of all the eigenvalues in the calculation of the signal-to-noise ratio in the limit $n \to \infty$ and $q \to 0$ when $N \to \infty$.

The eigenvalues of M are given by equation 7.12 as

$$\lambda_\alpha = 1 - 2q \sin^2 \frac{\pi\alpha}{2n}$$

with $\alpha = 0, \ldots, n-1$, and hence all go to 1 in the limit we are discussing.

Denoting by S_α the contribution to the signal from the eigenvalue λ_α:

$$S_\alpha = \langle h^p_\alpha \rangle_{+1} - \langle h^p_\alpha \rangle_{-1}$$

The corresponding eigenvectors are, for n large

$$v^\alpha_k \simeq \sqrt{\frac{2}{n}} \cos \frac{\pi k \alpha}{n-1}$$

Hence, for α even, S_α=0, and for α odd, with J_m given by equation 7.1,

$$S_\alpha \simeq \lambda^{p-1}_\alpha \frac{8q}{n^2} \sum_{m=0}^{n-1} J_m \cos\left(\frac{\pi\alpha m}{n-1}\right) \simeq \lambda^{p-1}_\alpha \frac{8q}{n} \int_0^1 (1-2x)\cos(\pi\alpha x)\, dx \quad \text{(B.1)}$$

when n is large. Carrying out the integration one has

$$S_\alpha \simeq \frac{32q}{n\pi^2} \frac{\lambda^{p-1}_\alpha}{\alpha^2}$$

for α an odd integer and S_α=0 otherwise.

The total signal is

$$S \simeq \frac{q}{n} \sum_{\text{odd } \alpha} \left(1 - 2q \sin^2 \frac{\pi\alpha}{2n}\right)^{p-1} \frac{32}{(\alpha\pi)^2} \quad \text{(B.2)}$$

The sum in the above expression for S neither diverges nor vanishes as $n \to \infty$. It cannot diverge since all the eigenvalues are less than 1, so the series in α converges to a finite value. It cannot vanish because, when $p \sim x^{-1}$ as in equation 7.11, there is at least one λ^p_α that tends to a constant different from zero. Furthermore, its sum cannot vanish by a cancellation, since the number of positive terms is greater ($q < 1$) or equal ($q = 1$) to the number of negative ones and for each negative term there is a positive one with a greater absolute value.

As a consequence the inclusion of all the eigenvalues leaves the dependence of the signal on the learning parameters unchanged in the limit of large n and small q, it has the form equation 7.7.

The noise around +1 is equal to the noise around −1. In the limit $n \to \infty$ and $q \to 0$, it can be written as

$$\begin{aligned} R^2 &= \frac{1}{N}\left[\sum_J J^2 \left[\rho^p_J(1,1) + \rho^p_J(1,-1)\right] - \mathcal{O}\left(\frac{q^2}{n^2}\right)\right] \\ &= \frac{1}{N}\left[\sum_J J^2 \sum_K \left[\rho^1_K(1,1) + \rho^1_K(1,-1)\right] M^{p-1}_{KJ}\right] \end{aligned} \quad \text{(B.3)}$$

The second term on the right hand side of the first equality is the subtraction of $\langle h^p_i \rangle^2_\xi$ in equation 2.4. The sum of the two conditional distributions ρ^1, given by equation 7.6, gives a uniform vector proportional to the asymptotic distribution. So the vector $\rho^1_K(1,1) + \rho^1_K(1,-1)$ is invariant when multiplied by matrix M and hence

$$R^2 = \frac{1}{N}\left[\sum_{m=0}^{n-1} J^2_m \rho^\infty_{J_m}\right]$$

So the asymptotic behavior of the noise, as a function of n and q, preserves its form in the case of a single dominant eigenvalue, equation 7.8.

References

Amaldi, E., and Nicolis, S. 1989. Stability-capacity diagram of a neural network with Ising bonds. *J. Phys. France* **50**, 2333.

Amit, D. J., and Brunel, N. 1993. Adequate input for learning in attractor neural network. *NETWORK* **4**, 177.

Amit, D. J., and Fusi, S. 1992. Constraints on Learning in Dynamic Synapses. *NETWORK* **3**, 443.

Amit D. J., Fusi, S., Genovese, S., Badoni, D., Riccardi, R., and Salina, G. 1992. LANN: Learning attractor neural network, model and hardware implementation, (INFN internal report, in Italian).

Amit, D. J., Gutfreund, H., and Sompolinsky, H. 1987. Statistical mechanics of neural networks near saturation. *Ann. Phys.* **173**, 30.

Badoni, D., Riccardi, R., and Salina, G. 1992. Learning attractor neural network: The electronic implementation. *International Journal of Neural Systems*, Vol. 3, pp. 13–24.

Brunel, N. 1993. Private communication.

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

Cox, D. R., and Miller, H. D. 1965. *Theory of Stochastic Processes.* Methuen, London.

Griniasty, M., Tsodyks, M. V., and Amit, D. J. 1992. Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Comp.* **5**, 1–17.

Gutfreund, H., and Stein, Y. 1990. Capacity of neural networks with discrete couplings. *J. Phys A: Math. Gen.* **23**, 2613.

Hopfield, J. J. 1982. Neural networks and physical systems with emergent selective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554.

Krauth, W., and Mezard, M. 1989. Storage capacity of memory networks with binary couplings. *J. Phys. France* **50**, 3057

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

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

Parisi, G. 1986. A memory which forgets. *J. Phys.* **A19**, L617.

Stanton, P. K., and Sejnowsky, T. J. 1989. *Nature (London)* **339**, 215.

Tsodyks, M. 1990. Associative memory in neural networks with binary synapses. *Modern Phys. Lett.* **B4**, 713.

Weisbuch, and Fogelman-Souliè, F. 1985. Scaling laws for the attractors of Hopfield networks. *J. Phys. Lett.* **2**, 337.

Willshaw, D. 1969. Non-holographic associative memory. *Nature (London)* **222**, 960.

Received May 19, 1993; accepted December 15, 1993.