A neural network model of working memory exhibiting primacy and recency

1991 J. Phys. A: Math. Gen. 24 1119

(http://iopscience.iop.org/0305-4470/24/5/025)

J. Phys. A: Math. Gen. 24 (1991) 1119–1135. Printed in the UK

A neural network model of working memory exhibiting primacy and recency

K Y M Wong†‡, P E Kahn† and D Sherrington†‡

† Department of Physics, Imperial College, London SW7 2BZ, UK
‡ Department of Theoretical Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK

Received 15 October 1990

Abstract. We consider a simple dilute neural network in which the synaptic strengths are bounded, and the probabilities of strengthening and weakening the synapses during learning are different. During the sequential learning of patterns, introvert networks (i.e. those with synapses more easily weakened than strengthened) exhibit recency (i.e. the preferential retention of the latest learned patterns) as in the Hopfield–Parisi model. On the other hand, extrovert networks (i.e. those with synapses more easily strengthened than weakened) exhibit both recency and the novel primacy effect (i.e. the preferential retention of the earliest learned patterns). The occurence of primacy depends on the initial distribution of the synaptic strengths. The relevance of the model to psychological experiments on working memory is also discussed.

1. Introduction

The Hopfield model [1] of neural networks symbolizes the initial success of applying statistical physics to models of associative memory [2]. Although this model is obviously far from realistic, it does reproduce qualitative features of biological neural networks such as stable memory within basins of attraction, error tolerance and robustness against noise.

Subsequent improvements to the model have often focused on introducing biologically motivated modifications, and have usually resulted in more realistic models of associative memory. Noting that, for instance, the synaptic strengths cannot be arbitrarily large, there have been proposals of models with either bounded synaptic strengths [3, 4] or with synaptic strengths renormalized during learning [4]. These models improve upon the original Hopfield model in a very important aspect: during sequential learning of patterns, the old patterns are automatically forgotten while only the most recent ones are recalled, avoiding the state of total confusion when too many patterns are stored in the original Hopfield model (the so-called 'memory catastrophe').

Another development in neural network modelling has been to take into account that in real systems the synaptic strengths are often asymmetric (i.e. $J_{ij} \neq J_{ji}$) and their connections are diluted [5]. This is in contrast to the original Hopfield model in which the synapses are all symmetric and fully connected. Extreme dilution, although it is not obviously in itself more realistic, simplifies the dynamical equations and allows the straightforward study of dynamical as well as equilibrium effects. Strikingly, however, the above-mentioned features of the Hopfield model are *still* retained.

In this paper we shall introduce, in addition, another modification to neural network modelling. From a physiological viewpoint, it is implausible that the mechanisms for the strengthening and weakening of synaptic strengths are the same [6]. Thus we shall study a dilute asymmetric network in which the bounded synaptic strengths are stochastically modified during learning, but the probabilities of strengthening and weakening are different. This is in contrast to the Hopfield–Parisi model [3] in which the modification of synaptic strengths (within bounds) is $\Delta J_{ij} \sim \xi_i^\mu \xi_j^\mu$ for the learning of the μth pattern, irrespective of whether ΔJ_{ij} strengthens or weakens the original J_{ij} (i.e. whether ΔJ_{ij} has the same or different sign as J_{ij}). When the synapses are more easily strengthened than weakened we can intuitively call the system a ***stubborn*** or ***extrovert*** network; otherwise it is termed ***stupid*** or ***introvert***; and that with equal likeliness for strengthening and weakening we call ***normal***. To avoid linguistic ambiguities we shall call the three cases extrovert, introvert and normal, which describe their tendencies towards stronger, weaker or normal synaptic strengths respectively. In order to focus on the phenomenological aspects of the model, we simplify the mathematical analysis by restricting the synaptic strengths to take the values ±1 and 0. Naively simple as it seems, the model already shows interesting features which have phenomenological relevance. These features are also present in models with bounded synapses taking a wider, or even continuous, range of values [7].

The phenomenology concerned in this model is the study in experimental psychology of working or short-term memory. The term working memory implies a system with limited capacity, for the temporary holding and manipulation of information during the performance of a range of cognitive tasks such as learning and retrieval [8]. Particularly relevant are experiments in which the subject is required to recall a list of items presented serially [9]. It is found that the probability of recalling an item depends on the serial presentation position. The increased probability of recall for the earliest learned items is called the primacy effect, whereas the recency effect corresponds to the increased probability of recall for the latest learned items.

Starting from *tabula rasa* (i.e. all the synaptic strengths are initialized to zero), we shall demonstrate that introvert networks exhibit recency but not primacy effects, which is qualitatively similar to the Hopfield–Parisi model. On the other hand, extrovert networks behave differently. The *tabula rasa* initial condition ensures that the earliest patterns are always learned by the strengthening of the synapses, since no previous information is present. Later patterns are then less well embedded in the network because they may require changes in the direction that weakens the synapses, which are more difficult in extrovert networks. Thus, on top of the usual forgetting mechanism due to pattern interference, extrovertness contributes another factor towards the deterioration of memory span when patterns are learned sequentially. As we shall see, these two factors are responsible for the presence of both primacy and recency effects in extrovert networks. We shall also demonstrate the inherent limited capacity of our three-state model, with patterns stored only temporarily as learning proceeds. Our model is thus a plausible candidate for a working memory, although its relationship to experimental psychology is far from straightforward, as is later brought out.

It is important to stress, however, the dependence of the primacy effect on the initial configuration of the synaptic strengths. A pre-requisite for the presence of primacy is that the initial distribution of synaptic strengths should be different from the asymptotic distribution after a large number of patterns have been learned. Furthermore, the initial distribution should be such as to enable a pattern to be better embedded in the network as compared with embedding it in the asymptotic distribution. This implies that for an extrovert network to exhibit primacy it must have an initial distribution weighted more towards the centre than the asymptotic distribution. In particular for the three-state model we study here, this requires that a sufficiently large fraction of the synaptic strengths should be initialized to zero. This dependence on the initial configuration is also studied in this work.

2. The model

The model we consider has a structure analagous to the diluted asymmetric Hopfield model introduced by Derrida *et al* [5], but with differently chosen synaptic weights. It consists of a network of N binary neurons $\sigma_i = \pm 1$ $(i = 1, \ldots, N)$ with synaptic interactions chosen as follows. Independently for each (ij) permutation, a synapse from j to i is present with a small probability, C/N, so that the average number of neurons feeding any other is C. Equivalently, the synaptic efficacy from j to i is given by $\tilde{J}_{ij} = J_{ij}C_{ij}$, where the $C_{ij} \in \{0, 1\}$ are independent parameters chosen at random according to the distribution

$$\rho(C_{ij}) = \frac{C}{N}\delta(C_{ij} - 1) + (1 - C/N)\delta(C_{ij}) \ . \tag{1}$$

C is chosen so that $C \ll \ln N$ and the limit $N \to \infty$ is taken in the analysis. This restriction on the average number of connections ensures that in any finite number of steps of the dynamics, any neuron almost never receive inputs from itself, even indirectly. Correlations between the neurons are thus eliminated, and it is this simplification that allows for an exact solution of the dynamics.

The synapses that are present are restricted to take one of the values ±1 or 0. Starting from *tabula rasa* (i.e. all J_{ij} are initialized to zero), random patterns $\{\xi_i^\mu\}_{i=1}^N$ (where ξ_i^μ denotes the value of neuron i for pattern μ) are learned at the rate of one pattern per unit learning time. Since the synapses can only take a finite number of values, the system can only store a few patterns if all the synapses are updated during the learning of a pattern. Thus each pattern is learned by randomly choosing only a fraction f of the synapses, and stochastically updating the chosen synaptic strengths J_{ij} by an increment $\xi_i^\mu\xi_j^\mu = \pm 1$, except when such an increment is not allowed by the synaptic bounds. The probabilities of updating are p_{r} and p_{c} if the increment $\xi_i^\mu\xi_j^\mu$ has the same or opposite sign as the synaptic strength J_{ij} respectively. Thus p_{r} and p_{c} are the registration (or strengthening) and correction (or weakening) probabilities. In the present three-state model, p_{c} effectively adjusts the likelihood of change from the synaptic values ±1, while p_{r} modifies change from 0. If $p_{\mathrm{c}}/p_{\mathrm{r}} < 1$ the network is termed extrovert, with synapses resisting change from the values ±1, if $p_{\mathrm{c}}/p_{\mathrm{r}} > 1$ it is termed introvert, with synapses biased towards 0 and if $p_{\mathrm{c}}/p_{\mathrm{r}} = 1$ it is referred to as normal. Each pattern of a particular sequence of patterns is then learned in turn in this manner, thus determining the values of the synapses at any time. In fact, as we show below, the properties of the network depend on $(f, p_{\mathrm{c}}, p_{\mathrm{r}})$ only in the combinations $fp_{\mathrm{r}}, fp_{\mathrm{c}}$. We find it convenient conceptually, however, to think in terms of $p_{\mathrm{c}}, p_{\mathrm{r}} \sim \mathrm{O}(1)$ with any c-dependence contained in f.

It is convenient to study the retrieval properties of this system by considering the probability distribution of the synaptic strengths. Let $P_+^\mu, P_0^\mu, P_-^\mu$ be respectively the probabilities of J_{ij} being in the states +1, 0, −1 after μ patterns have been learned. The probabilities before and after the learning of the μth pattern are related, when $\xi_i^\mu \xi_j^\mu = \pm 1$ respectively, by the matrix equation

$$P(\mu)_{ij} = \left[(1-f)\mathbf{I} + f\mathbf{T}_\pm\right] P(\mu-1)_{ij} \tag{2}$$

where the matrices $\mathbf{T}_\pm$ are given by

$$\mathbf{T}_+ = \begin{pmatrix} 1 & p_\mathrm{r} & 0 \\ 0 & 1-p_\mathrm{r} & p_\mathrm{c} \\ 0 & 0 & 1-p_\mathrm{c} \end{pmatrix} \qquad \mathbf{T}_- = \begin{pmatrix} 1-p_\mathrm{c} & 0 & 0 \\ p_\mathrm{c} & 1-p_\mathrm{r} & 0 \\ 0 & p_\mathrm{r} & 1 \end{pmatrix}$$

and $P(\mu)_{ij}$ is the transpose of $(P_+^\mu, P_0^\mu, P_-^\mu)$.

For the retrieval stage we consider either synchronous or asynchronous dynamics for the network [5], taking the post-synaptic potential thresholds to be zero and ignoring synaptic noise, so that at retrieval time t an updating event is defined by

$$\sigma_i(t+1) = \mathrm{sign}(h_i(t)) \qquad \text{synchronous} \tag{3a}$$

or probabilistically with updating occuring with frequency $\Delta\tau^{-1}$ according to

$$\sigma_i(t+\Delta\tau) = \mathrm{sign}(h_i(t)) \qquad \text{asynchronous} \tag{3b}$$

where $h_i(t)$ is the local field, or post-synaptic potential, on neuron i and is given by

$$h_i(t) = \sum_{j=1}^{N} J_{ij}\sigma_j(t)\ . \tag{4}$$

We are mainly interested in the asymptotic retrieval behaviour of the system, so both synchronous and asynchronous dynamics would give the same result in the present model.

We consider the time evolution of a state of the network having a macroscopic overlap m with a specified pattern μ

$$m_\mu(t) = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \sigma_i(t) \tag{5}$$

and a microscopic overlap with all other stored patterns. As shown by Derrida and Nadal [10], $m(t)$ asymptotically approaches the fixed point m^* of the equation

$$m_\mu^* = g(m_\mu^*) \tag{6}$$

where g is given in the limit of $C \to \infty$ (but still with $C \ll \ln N$) by

$$g(m) = \mathrm{erf}\left(\frac{m}{\sqrt{2}\Delta_\mu}\right) \tag{7}$$

where Δ_μ is the noise-to-signal ratio

$$\Delta_\mu = \left(\frac{(D_\mu - A_\mu^2)}{CA_\mu^2}\right)^{1/2} \tag{8}$$

with A_μ and D_μ given by the following averages over the randomly stored patterns:

$$A_\mu = \langle \xi_i^\mu \xi_j^\mu J_{ij} \rangle \qquad D_\mu = \langle J_{ij}^2 \rangle\ . \tag{9}$$

Determination of the storage properties of the network then follows from an analysis of equation (6).

3. Solution

3.1. Derivation of fixed point equation

Consider the retrieval properties of the $(r+1)$th pattern when s further patterns have been learned (i.e. at the learning time $T = r + 1 + s$). We proceed to calculate the probability distribution of the J_{ij} for such a sequence of patterns, and hence determine equation (6), by considering the average effects of the first r and last s patterns, and the explicit effect of the marked $(r+1)$th pattern. These averages are applicable as the strong dilution present in the model ensures that correlations between the J_{ij} can effectively be ignored.

Training the first r patterns starting from *tabula rasa*, and averaging over the random choice of those patterns, gives the probability distribution of the J_{ij} from equation (2) as

$$\boldsymbol{P}_{ij}^{(r)} = \left[(1-f)\mathbf{I} + f\left(\frac{\mathbf{T}_+ + \mathbf{T}_-}{2}\right)\right]^r \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \mathbf{T}^r \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} . \tag{10}$$

Here $\mathbf{T}$ is defined as a 3×3 matrix relating the probability distributions at adjacent stages of the training procedure. Thus $\mathbf{P}_{ij}^{(r)}$ can be expressed as a linear combination of the (unnormalized) eigenvectors $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ of $\mathbf{T}$, where

$$\mathbf{e}_1 = \frac{1}{2+R}\begin{pmatrix} 1 \\ R \\ 1 \end{pmatrix} \qquad \mathbf{e}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \qquad \mathbf{e}_3 = \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

with eigenvalues of $1, 1-\frac{1}{2}RE$ and $1-\frac{1}{2}(2+R)E$ respectively, $R \equiv p_{\mathrm{c}}/p_{\mathrm{r}}$ and $E \boxed{\equiv} fp_{\mathrm{r}}$. As we shall see, the $\mathbf{e}_2$ mode is associated with the evolution of the signal strength of the marked pattern, and the $\mathbf{e}_3$ mode with the evolution of the synaptic weight distribution due to the unmarked patterns. We shall call E the learning intensity, since it measures the intensity with which patterns are stored in the direction of strengthening the synapses. Equivalently, the network behaviour depends on both the registration and correction intensities $E_{\mathrm{r}} \equiv fp_{\mathrm{r}}$ and $E_{\mathrm{c}} \equiv fp_{\mathrm{c}}$ respectively; here we have chosen $E \equiv E_{\mathrm{r}}$ and $R \equiv E_{\mathrm{c}}/E_{\mathrm{r}} = p_{\mathrm{c}}/p_{\mathrm{r}}$ as independent variables to facilitate discussion. Furthermore, we shall consider $fp_{\mathrm{c}}, fp_{\mathrm{r}} \sim \mathrm{O}(1/\sqrt{C})$, for reasons to become evident later.

For sufficiently small $fp_{\mathrm{c}}, fp_{\mathrm{r}}$, we may easily evaluate $\boldsymbol{P}_{ij}^{(r)}$ by using the approximation $(1+x)^r = \exp(rx)$. Then introducing the $(r+1)$th pattern into the probability distribution explicitly, and considering the effect of the next s patterns in a similar way to the first r patterns, we find that

$$\begin{aligned} \boldsymbol{P}_{ij}^{(r+1+s)} &= \mathbf{e}_1 + \xi_i^{r+1}\xi_j^{r+1}\frac{E}{2(2+R)}\{2R + (2-R)\exp[-\tfrac{1}{2}(2+R)Er]\}\exp(-\tfrac{1}{2}REs)\mathbf{e}_2 \\ &\quad + \left(\frac{1}{(2+R)} - \frac{E}{2}\right)\exp[-\tfrac{1}{2}(2+R)E(r+s)]\mathbf{e}_3 . \end{aligned} \tag{11}$$

The relevant averages to determine equation (6) for $\mu = r+1$ are then given by

$$A_{r+1} = \xi_i^{r+1}\xi_j^{r+1}\mathbf{e}_2 \cdot \boldsymbol{P}_{ij}^{(r+1+s)} = \frac{E}{2+R}\{2R+(2-R)\exp[-\tfrac{1}{2}(2+R)Er]\}\exp(-\tfrac{1}{2}REs) \tag{12}$$

$$D_{r+1} = (1\ 0\ 1)\cdot \boldsymbol{P}_{ij}^{(r+1+s)} = \frac{2}{2+R}\{1 - \exp[-\tfrac{1}{2}(2+R)E(r+s)]\} + \mathrm{O}(E) \tag{13}$$

and the noise-to-signal ratio in (8) becomes

$$\Delta_{r+1} = \frac{\sqrt{2(2+R)\{1-\exp[-\frac{1}{2}(2+R)E(r+s)]\}}}{\sqrt{C}E\{2R+(2-R)\exp[-\frac{1}{2}(2+R)Er]\}\exp(-\frac{1}{2}REs)} \, . \tag{14}$$

Substituting Δ_{r+1} into equation (6) gives the fixed point equation which determines the retrieval behaviour of the network.

Before beginning the formal analysis it is worth noting the terms that will determine the broad nature of the network's storage properties. We see that the noise-to-signal ratio is dependent on three factors. The signal term A_{r+1} of the marked pattern decays exponentially with s on a 'time' scale of $2/RE$, since it is associated with the $\mathbf{e}_2$ mode. This decay is due to the interfering effect of the subsequently stored patterns when the synaptic strengths are bounded. (Note that the signal term will not decay when the synapses are unbounded; all the stored patterns are simply confused when the storage capacity is reached, constituting the so-called 'memory catastrophe'.)

Secondly, the pre-factor of the signal term, $\{2R+(2-R)\exp[-\frac{1}{2}(2+R)Er]\}/(2+R)$, reflects how favourable the synaptic distribution, at the instant immediately before the marked pattern is learned, is for the marked pattern to be embedded. Its value evolves from 1 to $2R/(2+R)$ on a 'time' scale of $2/(2+R)E$ as learning proceeds. For extrovert networks ($R < 1$), this means that earlier learned patterns are better embedded than later ones, because earlier learned patterns result in more synaptic updating when the synaptic distribution is weighted more towards the centre.

Finally, the noise term D_{r+1} grows with the total number of stored patterns. Its value increases from 0 to its maximum value, again on a time scale of $2/(2+R)E$ associated with the $\mathbf{e}_3$ mode. It is the non-trivial interplay of these three factors that gives rise to the novel effects that are outlined below.

3.2. Analysis of fixed point equation

We shall first consider the memory lifetime $L(r)$ when r patterns have previously been presented (i.e. of the $(r+1)$th) pattern. This is given by the value of s when the pattern is just forgotten (no longer retrievable), and hence by considering the solutions of equation (6) in the limit $m^*_{r+1} \to 0$ (i.e. the solutions of $g'(0) = 1$). Thus $L(r)$ is given by the value of s for which $\Delta_{r+1} = \sqrt{2/\pi}$. This is equivalent to the equation

$$\frac{E}{E^*} = \frac{\sqrt{1-\exp[-\frac{1}{2}(2+R)E(r+L(r))]}}{\{1+[(2-R)/2R]\exp[-\frac{1}{2}(2+R)Er]\}\exp(-\frac{1}{2}REL(r))} \tag{15}$$

where

$$E^* = \frac{\sqrt{2+R}}{2R}\sqrt{\frac{\pi}{C}} \, .$$

In the limit of $r \to \infty$, that is once the memory has been saturated, the $(r+1)$th pattern is forgotten after an additional $L(\infty)$ patterns have been stored, where

$$L(\infty) = \frac{2}{RE}\ln\left(\frac{E}{E^*}\right) \, . \tag{16}$$

$L(\infty)$ is thus the asymptotic storage capacity of the network. It reaches its maximum at $E = eE^*$, and the maximum storage capacity is $O(\sqrt{C})$.

This storage capacity is clearly drastically reduced in comparison to the more usual value of $O(C)$ for other dilute networks. This is not simply a consequence of the synapses taking only three values, but also of the sequential nature of the training procedure for storing patterns. For example Sompolinsky [11] has considered a comparable recursive network with fully connected synapses. In it the synapses are unrestricted until they have been fully trained, after which they are clipped to the values $\pm 1, 0$. The capacity is then of the same order as a model without such restriction on the magnitude of the synapses. It does, however, suffer from the same total collapse of memory beyond the critical capacity. What is surprising about our model, however, is that the retrievable storage capacity is as high as it is, given the degree to which information about earlier patterns is lost as new patterns are stored. Furthermore, such limited capacity need not be regarded as restrictive as it is an integral part of any model of working memory.

If we choose the intensity as a higher power of C than $C^{-1/2}$, the capacity $L(\infty)$ is then smaller. For example $E \sim O(1)$, corresponding to $f \sim O(1)$ for $p_r, p_c \sim O(1)$, would imply $L(\infty) \sim O(\ln C)$. This dependence on the learning intensity reflects further the difference between sequential presentation to a bounded synapse system and the effects of eventual clipping of synapses after presentation to an unbounded system.

In the limit of $r \to 0$ equation (15) reduces to

$$\frac{E}{E^*} = \frac{\sqrt{1 - \exp[-\frac{1}{2}(2+R)EL(0)]}}{[(2+R)/2R]\exp(-\frac{1}{2}REL(0))} \tag{17}$$

which determines the lifetime $L(0)$ of the first learned pattern. Figure 1 plots the rescaled lifetimes $\lambda \equiv L/\sqrt{C}$ of the first pattern and of any pattern stored after the asymptotic capacity has been reached, against the rescaled intensity $\epsilon \equiv E\sqrt{C}$ for a normal ($R = 1$) network. Clearly there is a critical value of the intensity, $\epsilon^* = \sqrt{(2+R)\pi/2R}$, below which the network fails to store any more than the first few patterns, which are themselves lost as training continues. A similar critical value of the intensity with which patterns are stored is seen in the work of Mézard *et al* [5]. Note that the lifetimes first increase with ϵ, reach their maxima and then decrease. Thus, for sufficiently small E, the patterns are better embedded as the intensity increases, but pattern interference degrades the lifetimes when the intensity is excessive.

3.3. Above threshold intensity

We first consider the case $\epsilon > \epsilon^*$. For a fixed value of ϵ, we plot in figure 2 the lifetime of each pattern in the sequence as a function of the rescaled learning position $\rho \equiv r/\sqrt{C}$ for extrovert, normal and introvert networks (extremely introvert networks, that is with $R > 2$, are discussed separately in section 3.5).

We see, for the cases considered here, that patterns learned at the start are remembered for longer than patterns learned later on, whose lifetime monotonically decreases to a constant value. This is due to the increasing pattern interference as learning proceeds, and is also observed in previous models [4, 10]. What is of particular

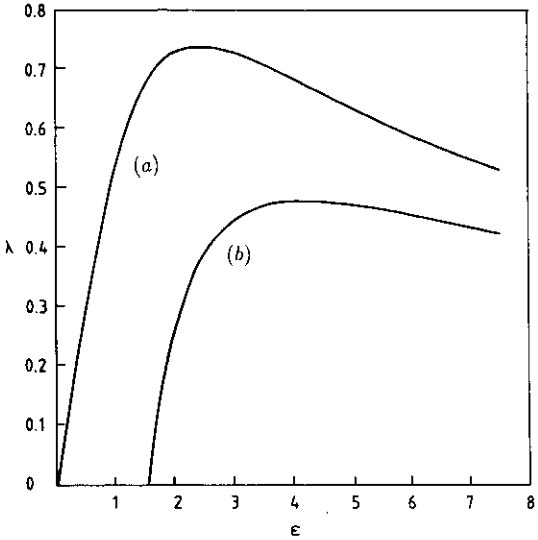

Figure 1. Curve a gives the lifetime $\lambda \equiv L/\sqrt{C}$ of the first stored pattern as a function of ϵ, the intensity of storage, for $R = 1$. The lower curve b gives the corresponding lifetime for a pattern stored after the memory has been saturated.

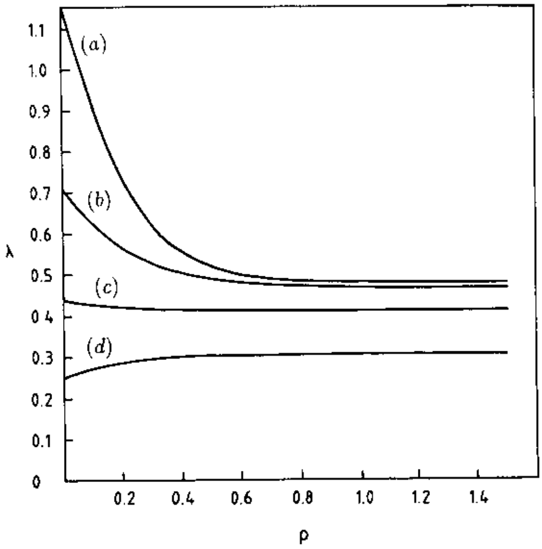

Figure 2. The lifetime $\lambda \equiv L/\sqrt{C}$ for $R = 0.5, 1, 2$ and 5 (curves a, b, c and d respectively) is given as a function of $\rho \equiv \tau/\sqrt{C}$ at $\epsilon = \epsilon_{\mathrm{obs}}(R) \equiv (0.75 + 0.5R^{-1})e\sqrt{\pi/(2+R)}$, which is arbitrarily chosen to ensure that the novel effects are easily observable.

interest, however, is the rate of this decrease in pattern lifetime. We see by partially differentiating equation (15) with respect to ρ, and evaluating at $\rho = 0$, that

$$\left.\frac{\partial\lambda}{\partial\rho}\right|_{\rho=0} = -\frac{2(2-R) + (3R-2)\exp[-\frac{1}{2}(2+R)\epsilon\lambda]}{2R + (2-R)\exp[-\frac{1}{2}(2+R)\epsilon\lambda]} \; . \tag{18}$$

Clearly if $R < 1$ we see that $\partial\lambda/\partial\rho|_0 < -1$. Thus for an extrovert network the rate at which the pattern lifetime decreases is faster than the rate at which new patterns are stored, until we reach pattern ρ_0, where $\partial\lambda/\partial\rho|_{\rho_0} = -1$. Here ρ_0 is given by

$$\rho_0 = \frac{2}{(2+R)\epsilon} \ln\left(\frac{2-R}{R^2}\right) . \tag{19}$$

This ensures that, for a certain period within the training procedure, the patterns at the start of the sequence may be stored for longer, and retrieved with a higher output overlap than patterns in the middle, some of which may be entirely forgotten.

The extent of this effect, which we term the primacy effect, and its absence from a normal or an extrovert network, is evident in figure 3. This indicates the stages of the training procedure for which each pattern may be retrieved, as a patttern learned at the rescaled time ρ is retrievable in the learning time interval τ enclosed between the curves $\tau = \rho$ and $\tau = \rho + \lambda(\rho)$. The positions ρ of the patterns still memorized by the network at a learning time, say x, are determined by the segment of the horizontal line $\tau = x$ enclosed in the area bounded by the two curves.

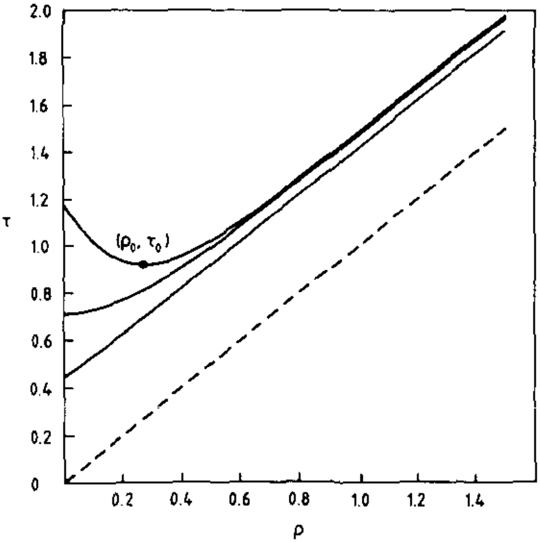

Figure 3. The curves $\tau = \rho + \lambda(\rho)$ for $R = 0.5, 1$ and 2 (a, b and c respectively), at $\epsilon = \epsilon_{\rm obs}(R)$ and the line $\tau = \rho$, plotted against ρ.

As indicated by curves b and c for normal and introvert networks ($R \geq 1$), all the stored patterns are retrievable for $\tau \leq \lambda(0)$. The forgetting of patterns, starting from the earliest learned ones, takes place when $\tau > \lambda(0)$. The latest patterns are recallable. We call this the recency effect, which is already observed in previous models [4, 10]. No primacy effect is present.

For extrovert networks ($R < 1$), curve a shows that all the stored patterns are retrievable up to the learning time $\tau_0 \equiv \rho_0 + \lambda(\rho_0)$. As learning proceeds further, patterns in the middle positions around ρ_0 start to be forgotten. This continues until the learning time $\lambda(0)$, when the first pattern is forgotten. Beyond $\lambda(0)$, only the latest learned patterns are retrievable. The preferential retention of memory of the earliest learned patterns between τ_0 and $\lambda(0)$ is the primacy effect.

Even before forgetting sets in at the learning time τ_0, the primacy phenomenon is already present in the quality of retrieval as a function of patterns presented from the start of training, and persists up to the learning time $\lambda(0)$, when the first pattern is forgotten and the primacy effect vanishes. These characteristics are indicated in figure 4, which gives the retrieval quality of each pattern at several stages in the training procedure.

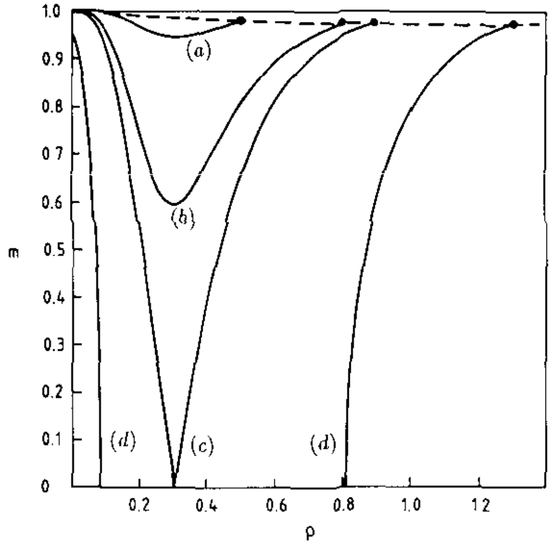

Figure 4. The retrieval overlap m as a function of $\rho \equiv r/\sqrt{C}$ at $\epsilon = \epsilon_{\rm obs}(R)$ and for $R = 0.25$, at a stage in the training procedure (a, b) between $\tau = \rho_0$ and $\tau = \tau_0 = \rho_0 + \lambda(\rho_0)$, (c) at $\tau = \tau_0$ and (d) between $\tau = \tau_0$ and $\tau = \lambda(0)$, respectively. The end points of each curve indicate the learning position of the last learned pattern; their envelope is given by the monotonically decreasing dotted line.

We note that the primacy effect is dominant for learning times τ below ρ_0, so that the retrieval quality of a pattern is monotonically decreasing with its position ρ. Recency effects set in, for τ greater than ρ_0, when the retrieval quality monotonically increases thereafter. The pattern position for this onset of recency improvement is independent of τ, and is always at ρ_0.

Finally we derive the rescaled storage capacity $\alpha \equiv p/\sqrt{C}$, where p is the number of retrievable patterns stored in the network, as a function of the learning time. For an extrovert network there are three regions.

(i) $0 < \tau < \tau_0$; each stored pattern can be retrieved until τ_0, when forgetting first starts, and hence $\alpha = \tau$.

(ii) $\tau_0 < \tau < \lambda(0)$; between these times the patterns from positions $\rho_<$ to $\rho_>$ are forgotten, where $\rho_<$ and $\rho_>$ are the two solutions of $\tau = \rho + \lambda(\rho)$. Hence $\alpha = \tau - \rho_> + \rho_<$. This is the region where both primacy and recency effects are present.

(iii) $\lambda(0) < \tau$; after $\lambda(0)$ only the recency effect is present with ρ_1 patterns forgotten, where ρ_1 is the single solution of $\tau = \rho + \lambda(\rho)$, and all later patterns retrievable. Hence $\alpha = \tau - \rho_1$.

The storage capacity α is plotted as curve a of figure 5; note the presence of three distinct regions. Figure 5 also shows the corresponding storage capacity for the normal and introvert networks, for which only two regions exist as no primacy effect is present.

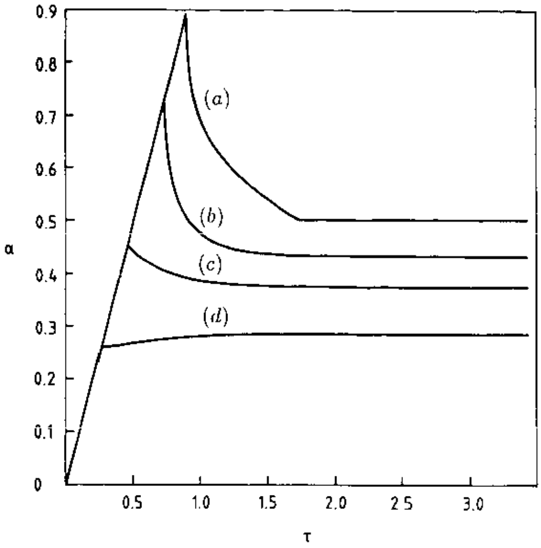

Figure 5. The storage capacity $\alpha \equiv p/\sqrt{C}$, as a function of τ, at $\epsilon = \epsilon_{\text{obs}}(R)$, for $R = 0.25, 1, 2$ and 5 (curves a, b, c and d respectively).

Note the similarity of these curves b and c in figure 5 with those of [5] and [10].

3.4. Below threshold intensity

Next we discuss the case of an extrovert network when $\epsilon < \epsilon^*$. In this case patterns can be learned at the initial stage, but all of them will be forgotten as learning proceeds, until eventually no patterns can be learned by the system. Hence no extended memory is possible. However, even in this regime, an extrovert network shows interesting behaviour, as demonstrated in figure 6(b)–(d); curve a is for $\epsilon > \epsilon^*$ and is for comparison.

For ϵ just below ϵ^* (curve b), the network exhibits both primacy and recency effects, although the recency effect in this case is restricted, and patterns cannot be learned indefinitely. As learning proceeds, the patterns in some middle position are the first ones to be forgotten, and the latest learned ones are the last to be forgotten. We call this a stronger recency (SR) regime, as the recency effect persists longer than the primacy effect.

For lower values of ϵ (curve c), the network continues to exhibit both primacy and (restricted) recency effects. Forgetting still starts in some middle position, but the earliest learned patterns are the last ones to be forgotten. We thus call this a stronger primacy (SP) regime, as the primacy effect now persists longer than recency. For still lower values of ϵ (curve d), the network exhibits primacy effect only (PO). Forgetting starts from the latest learned patterns, until eventually the earliest ones are forgotten. These three regimes for different values of ϵ, along with the region of extended memory (EM) for $\epsilon > \epsilon^*$, are shown in figure 7.

In the limit of extremely small ϵ, patterns are still faintly memorized in the diluted model. More interestingly, the slope $\partial\lambda/\partial\rho$ approaches -1 for all the memorized patterns. This implies that all the patterns are forgotten almost simultaneously at the same learning time, thus recovering the situation of memory catastrophe reminiscent of the Hopfield network without synaptic bounds. Indeed, when ϵ is small, updating

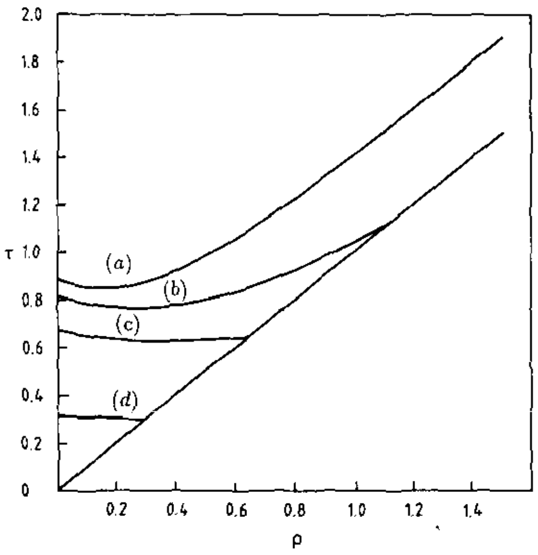

Figure 6. The curves $\tau = \rho + \lambda(\rho)$ for $\epsilon = 3.0, 1.7, 1.2$ and 0.5 (curves a, b, c and d respectively), with values of ϵ chosen so as to highlight the corresponding regimes of extended memory (EM), stronger recency (SR), stronger primacy (SP), and primacy only (PO)) when $R = 0.8$, and the line $\tau = \rho$, plotted against ρ.

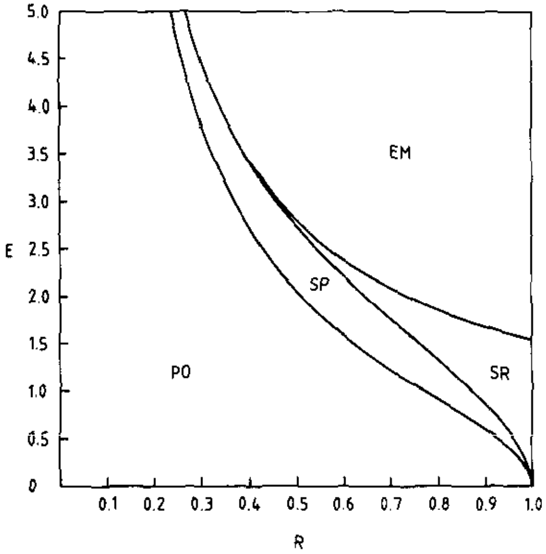

Figure 7. *Memorv behaviour in the space of the ratio* R and intensity ϵ, showing the regimes of extended memory (EM), stronger recency (SR), stronger primacy (SP) *and primacy only* (PO).

events for each synapse become very sparse during the period when learning takes place, and the effects of synaptic bounds become vanishing. This causes the learning behaviour of the network to approach that of the unbounded network.

3.5. Extremely extrovert networks

The case of very extrovert networks deserves special attention. When $R << 1$, equation (15) can be modified into

$$\frac{\epsilon}{\sqrt{\pi/2}} = \frac{\sqrt{1-\exp[-\epsilon(\rho+\lambda)]}}{\exp(-\epsilon\rho - \frac{1}{2}R\epsilon\lambda)} \; . \tag{20}$$

As before, the system does not exhibit extended memory below the threshold intensity $\epsilon^* \sim \sqrt{\pi/2}/R$. However, the system has interesting behaviour even in this regime, provided that $\epsilon > \sqrt{\pi/2}$. Here the lifetime in the early and later learning stages behave differently. In the early learning stage, patterns are efficiently embedded, and the extreme resilience against weakening the synapses preserves the learned patterns in subsequent learning time. Thus the lifetimes are long and $\lambda \sim \mathrm{O}(R\epsilon)^{-1}$. Precisely,

$$\lambda = \frac{2}{R\epsilon}\left(\ln\left[\frac{\epsilon}{\sqrt{\pi/2}}\right] - \epsilon\rho\right) \; . \tag{21}$$

The lifetime decay is caused by the increasing ineffieciency in embedding later learned patterns, and has a slope of $-2/R$.

When $\rho \sim \epsilon^{-1}\ln(\epsilon/\sqrt{\pi/2})$, the network behaviour crosses over to a regime of short lifetime, where $\lambda \sim \mathrm{O}(\epsilon)^{-1}$. Here the synaptic weights are already fixed by the network's previous learning experience, and the extreme resistance to weaken the synapses prevents the efficient embedding of newly learned patterns. Thus

$$\lambda = -(1/\epsilon)\ln[\exp(\epsilon\rho) - (2\epsilon^2/\pi)\exp(-\epsilon\rho)] \; . \tag{22}$$

The lifetime decay has a slope of O(1), meaning that the patterns learned at this stage are forgotten in a nearly catastrophic way, leaving the earlier learned patterns with a much longer span. Patterns with positions $\rho \geq \rho(0)$ cannot be learned, where

$$\rho(0) = \epsilon^{-1}\ln\left[\frac{1}{2}\left(1+\sqrt{1+\frac{8\epsilon^2}{\pi}}\right)\right] \; . \tag{23}$$

When $R = 0$ exactly, patterns with positions $\rho \leq \rho(\infty)$ have infinite lifetimes, where

$$\rho(\infty) = \epsilon^{-1}\ln(\epsilon/\sqrt{\pi/2}) \tag{24}$$

and patterns with positions $\rho(\infty) < \rho < \rho(0)$ have finite lifetimes. This means that the network exhibits permanent primacy, similar to models of long-term memories [10, 12, 13]. The long lifetimes of the early learning stage discussed above can therefore be considered as a precursor of permanent primacy. Permanent primacy is previously observed in the nonlinear model of Burgess *et al* [14, 15]; the existence of the 'point of no return' therein is the equivalent of our absolute extrovertness $R = 0$.

3.6. Extremely introvert networks

Novel effects are also revealed for sufficiently introvert networks. We see in curve d of figure 2 that the pattern lifetime steadily increases as patterns are learned, towards its asymptotic value. The maximum storage capacity is thus no longer given when forgetting starts, but is instead effectively given when a large number of patterns have been stored, as indicated in the lowest curve of figure 5. A simple analysis of equation (15) reveals that this new behaviour is present only if $R > 2$ and the intensity ϵ is strong enough. Indeed, for an introvert network, information is better embedded when the synaptic distribution is weighted away from the centre. Starting from *tabula rasa*, this condition can only be attained at a later stage of the training procedure, thus accounting for the longer lifetimes of the latest learned patterns. This interpretation is apparent in the change of behaviour in the signal term in (12) when R becomes greater than 2.

4. The origin of the primacy and recency effects

As an alternative to starting the training procedure from *tabula rasa* we now consider starting from synapses with an arbitrary symmetric initial probability distribution having a fraction p_0 of zero synapses, with an aim to analysing how this new condition modifies the primacy and recency effects.

Clearly the asymptotic distribution of the synapses after a large number of patterns have been stored will remain the same whatever the starting conditions. It is also manifest that if we started from this asymptotic distribution no primacy effect would appear as this would be equivalent to storing patterns in the limit $r \rightarrow \infty$ of section 2, where no primacy effect is present. This would be the case no matter how extrovert the network is.

Performing an analysis similar to that of section 3, we find that the condition for the appearance of the primacy effect, namely that the rate at which the pattern lifetime decreases for the earlier patterns is faster than the rate at which new patterns are stored (i.e. $\partial\lambda/\partial\rho|_{\rho=0} < -1$), is given by

$$R \leq \frac{2p_0}{1+p_0} \tag{25}$$

thus ensuring the onset of the primacy effect only if the network is sufficiently extrovert. Alternatively, for a fixed value of R, the initial distribution p_0 for the onset of primacy has to be greater than $R/(2-R)$, which is in turn greater than the equilibrium distribution $R/(2+R)$.

Clearly the primacy effect is dependent on the proportion of synapses that are set to 0 at the start of training. If this is greater than the asymptotic proportion after a large number of patterns have been stored then, for a suitably extrovert network, the primacy effect is present. This may be accounted for by noting that in an extrovert network the synapses resist change from the values ±1 and thus the earliest learned patterns will determine the signs of the initially zero-valued synapses whose change are resisted during the learning of later patterns. These earliest patterns are thus remembered for longer than they would be otherwise. This, when combined with the decrease in pattern lifetime due to the noise induced as more patterns are stored, accounts for the origin of the primacy effect in our model.

The source of the recency effect is apparent in that the nature of the training procedure ensures that information about the earlier patterns is gradually lost, with the most recently learned patterns better embedded among the values of the synapses.

5. Conclusion

We have studied a model of working memory with variable registration and correction probabilities, and have found that extrovert networks (in which registration is stronger than correction) exhibit both primacy and recency effects during sequential learning of patterns, when the learning intensity is sufficiently strong (on a scale of order $1/\sqrt{C}$ where C is the connectivity) and the initial distribution of synaptic strengths is biased towards weakness compared with the asymptotic limit. On the other hand, only recency effects are present in normal and introvert networks (respectively with equal registration and correction probabilities and correction diminished). The exhibition of both primacy and recency effects for a learning session parallels psychological experiments on working memories.

When compared with our model, most of the previous models exhibit either memory catastrophe beyond storage capacity [2], or purely recency effects [3,4]. There have also been models in which purely primacy effects are present. An example is the 'irreversible bounds' (or 'absorbing bounds') model [10,12,13], in which the synaptic strengths stick to the bounds once they are reached. This model has a limited memory storage, in which learning of new patterns is impossible after a sufficient number of patterns have been learned sequentially, thus it serves as a plausible model for long-term memory. In fact, it is equivalent to the extremely extrovert limit ($R \to 0$) in our model. Another example is the nonlinear model by van Hemmen *et al* [16], in which the learning rule is defined by some nonlinear function ϕ, so that

$$J_{ij}^{(\mu)} = \phi(\epsilon\xi_i^\mu\xi_j^\mu + J_{ij}^{(\mu-1)}) \ . \tag{26}$$

In the case (c) considered in [16], only the primacy effect is present, and it can be shown that the particular form of the function ϕ has the property that patterns are better embedded in the strengthening than the weakening direction. It is apparent that *in these models, the primacy effect is related to the extrovertness inherent in the synaptic learning rules*. By varying the ratio $R = p_c/p_r$, our model incorporates these models as special cases and, furthermore, extends to models with both primacy and recency effects present. Sequential learning in general nonlinear models, in which both primacy and recency effects are present, are also currently being studied by Burgess *et al* [14,15].

Concerning the variation of the storage capacity during sequential learning, previous studies have consistently found that the storage capacity rises linearly with learning time, reaches a maximum and then approaches an asymptotic level (for extended memory) or drops to zero (for restricted memory). Our study has revealed a much wider variety of behaviour when the ratio R and the intensity E are varied. In the primacy regime we have found that the storage capacity has two kinks instead of one, one corresponding to the moment when forgetting starts, and the other to the moment when the earliest or the latest learned patterns are all forgotten. In the recency regime, we also found instances in which the storage capacity further increases after the moment when forgetting starts, and so we see that the kink does not always give the maximum storage capacity. By varying also the initial synaptic distribution, other behaviours are likely to be revealed.

We have also emphasized the relevance of a favourable initial synaptic distribution to the occurence of primacy. For symmetric distributions, we have shown that the initial distribution has to be comfortably distinct from the equilibrium distribution, in such a way that the earliest patterns are embedded in the synapses with sufficient advantages over the latter ones. In the three-state model, this means that the initial fraction of zero synapses for the occurence of primacy should not merely be greater than the equilibrium fraction $R/(2+R)$, but greater than $R/(2-R)$ as well.

The importance of favourable initial distributions distinct from the equilibrium distribution of synaptic strengths leads to an interesting issue. If we take a neural network which has undergone learning for a long time, its synaptic strengths will reach the equilibrium distribution, and the system will show no primacy behaviour in a new but equivalent learning exercise. Surely some discontinuity in the subsequent synaptic strength distribution is required to demarcate the starting of a new learning session. A possible mechanism for a favourable resetting of the distribution is that the synaptic strengths relax randomly to zero when learning stops. Another possibility is that the necessary discontinuity in the learning environment is inherent when a learning session consists of correlated patterns whose correlations are different from the previous ones. These ideas deserve further investigation.

Alternative explanations of the primacy effect exist. It is possible that the network has a high 'attention' at the beginning of a learning session, and thus the earliest learned patterns are embedded with a stronger intensity E. As learning proceeds, the attention is lowered and the intensity approaches an asymptotic value [17]. The relevance to psychological phenomena of this attention theory, as well as the extrovert theory proposed here and in [14, 15], has to be subject to comparison with differentiating psychological experiments. While both theories may eventually be relevant to psychological phenomena, it would be interesting to predict, for example, the consequences of presenting uncorrelated patterns among a majority of correlated ones during a learning session, while keeping the attention constant. Another current common explanation within experimental psychology for the primacy and recency effects is through an interaction of short- and long-term memory [9]. This interaction has been incorporated into a neural network that has been outlined and simulated by Schreter and Pfeifer in [18].

There are other more subtle aspects of the psychological phenomena of primacy and recency. One experimental result can be termed the quenching of primacy [9]. In this experiment, the subject is required to rehearse an item several times before the next one is presented. In this case, no primacy effect is present. The analogue of rehearsing in our model lies in increasing the learning intensity E. Roughly speaking, we observe similar consequences. For the extrovert network, we see that primacy effects are more dominant for low-intensity learning, whereas recency effects are more dominant for high-intensity learning. The temporal extent of the primacy effect is determined, roughly speaking, by the magnitude of the position ρ_0, which is inversely proportional to the intensity E according to equation (19). For sufficiently introvert networks, we even witness an increasing lifetime with pattern position for sufficiently high intensity, implying the opposite of primacy. However, we have not found any extrovert network which has both primacy and recency effects at low intensity but has primacy behaviour *completely* washed out at high intensity.

Another experiment can be termed the quenching of recency [9]. In this experiment, the subject is required to perform some interfering task immediately after the learning session. Afterwards, only primacy but no recency effect is present. If we assume that the interfering task corresponds to further presentation of random patterns in our model, this would correspond to the observation that out of all the patterns recallable up to a learning time ρ, only the first few are retrievable at a time later than ρ. This is clearly observed in comparing curves *b* and *d* in figure 4: the retrieval overlaps of patterns on curve *b* become 0 on the corresponding portion of curve *d*, except for those learned earliest, and surely recency is absent. However, psychology experiments imply that the pattern position ρ_0 of minimum overlap, i.e. the position for the onset of recency improvement, should shift in the increasing direction as learning proceeds, and this more subtle aspect is not observed in the present model.

Our model is of course idealized. The detailed updating mechanism need not be restricted to the stochastic Hebbian one we have chosen, nor need the synapses be discrete [7]. However, we believe that our simple study demonstrates that bounded synapses, with different probabilities for increase and decrease of their magnitudes on learning, are likely to be key ingredients in any self-contained model of short-term memory.

Acknowledgments

We thank J Shapiro, N Burgess and M Moore for many stimulating discussions, and for kindly communicating about their work prior to publication. We also thank A Treves, A Ho, L Henry and P Pinkava for helpful discussions. This work was supported by the SERC.

References

[1] Hopfield J J 1982 *Proc. Natl Acad. Sci. USA* **79** 2554

[2] Amit D J, Gutfreund H and Sompolinsky H 1987 *Ann. Phys., NY* **173** 30

[3] Parisi G 1986 *J. Phys. A: Math. Gen.* **19** L617

[4] Nadal J P, Toulouse G, Changeux J and Dehaene S 1986 *Europhys. Lett.* **1** 535

[5] Mézard M, Nadal J P and Toulouse G 1986 *J. Physique* **47** 1457
Derrida B, Gardner E J and Zippelius A 1987 *Europhys. Lett.* **4** 167

[6] Bear M F, Cooper L N and Ebner F 1987 *Science* **237** 42

[7] Kahn P E, Wong K Y M and Sherrington D 1990 *in preparation*

[8] Baddeley A 1986 *Working Memory* (Oxford: Clarendon)

[9] Atkinson R and Shiffrin R 1971 *Sci. Am.* **225** 82

[10] Derrida B and Nadal J P 1987 *J. Stat. Phys.* **49** 993

[11] Sompolinsky H 1986 *Phys. Rev.* A **34** 2571

[12] Peretto P private communication

[13] Gordon M B 1987 *J. Physique* **48** 2053

[14] Burgess N, Moore M A and Shapiro J L 1989 *Neural Networks and Spin Glasses, Proc. Porto Alegre* ed W K Theuman and R Koberle (Singapore: World Scientific)

[15] Burgess N, Shapiro J L and Moore M A 1990 *submitted to Network*

[16] van Hemmen J L, Keller G and Kuhn R 1988 *Europhys. Lett.* **5** 663

[17] Nadal J P, Toulouse G, Mézard M, Changeux J and Dehaene S 1987 *Computer Simulation in Brain Science* ed R M J Cotterill (Cambridge: Cambridge University Press) p221

[18] Schreter Z and Pfeifer R 1988 *Neural Networks from Models to Applications* ed L Personnaz and G Dreyfus (Paris: IDSET) p36