# Processing of Temporal Sequences in Neural Networks

H. Gutfreund[a] and M. Mezard
*Département de Physique de l'Ecole Normale Supérieure, F-75231 Paris, France*[b]
(Received 6 October 1987)

We study three ways of processing of temporal sequences of patterns stored in a neural network; retrieval, counting, and recognition. These three processes are studied analytically in a strongly diluted neural network. The results compare qualitatively to simulations in fully connected networks.

PACS numbers: 87.10.+e

The "standard" model of associative memory,[1,2] which has attracted much attention in the last few years, consists of a fully connected network of $N$ formal "neurons," represented by Ising spins $S_i$ $(i=1,\ldots,N)$. A set of $p$ uncorrelated patterns $\{\xi_i^\mu\}$ $(i=1,\ldots,N;\mu=1,\ldots,p)$, in which $\xi_i^\mu$ is either $+1$ or $-1$ with equal probability, is embedded in the interaction matrix, by the rule

$$J_{ij}^s=\frac{1}{N}\sum_\mu \xi_i^\mu \xi_j^\mu. \tag{1}$$

These patterns are the "memories" stored in the network. The retrieval of a particular memory is achieved when the system starting from some initial configuration (imposed by an external stimulus) evolves under its own dynamics to a stationary configuration $\{S_i\}$, which is strongly correlated with that memory. A fair amount of understanding of the properties of this model has been achieved by analytical analysis supplemented by numerical simulations.[3,4]

An important problem is to extend this model to networks which can associatively recall temporal sequences of patterns. Such an extension was already discussed in Refs. 1 and 2 and several proposals to achieve this goal have been made since then.[5-9] Previous treatments of sequences in networks with different architectures can be found elsewhere.[10-12]

Our discussion will be based on the model proposed in Refs. 6 and 7. There are now two sets of interactions between the neurons. One is the $J_{ij}^s$ of Eq. (1); its task is to stabilize the stored patterns. The second set of interactions, $J_{ij}^{tr}$, tends to induce transitions from one pattern to the next in a temporal sequence $\mu=1\to 2,\ldots,\to q$:

$$J_{ij}^{tr}=\frac{\lambda}{N}\sum_{\mu=1}^{q-1}\xi_i^{\mu+1}\xi_j^\mu. \tag{2}$$

This interaction acts with a certain time delay $\tau$. The dynamics is described by the time evolution of the overlaps:

$$\tilde{S}_\mu(t)=\frac{1}{N}\sum_i \xi_i^\mu S_i(t), \tag{3}$$

in terms of which the local field on spin $i$ reads

$$h_i(t)=\sum_{\mu=1}^{p}\xi_i^\mu \tilde{S}_\mu(t)+\lambda\sum_{\mu=1}^{q-1}\xi_i^{\mu+1}\tilde{S}_\mu(t-\tau). \tag{4}$$

The updating rule at finite temperature $T=1/\beta$ is

$$S_i(t+1)=\pm 1 \text{ with probability } (1+e^{\mp 2\beta h_i(t)})^{-1}. \tag{5}$$

One can distinguish between parallel dynamics, when all spins are updated simultaneously at discrete time steps, and sequential dynamics, when the spins are updated one at a time in a random order. Delay functions more general than (4) have been discussed in Ref. 6 and could be used here as well.

If the network has been in state $\nu$, namely $\tilde{S}_\nu\simeq 1$, for a time larger than $\tau$, and if $\lambda$ is sufficiently large, then the second term in (4) will cause a transition to state $\nu+1$. Thus, such a network can retrieve a temporal sequence in response to an external stimulus corresponding to the first pattern in the sequence.

It has been suggested that such a network can be used for the recognition of temporal sequences[7] and for counting the number of external signals, as for example chimes.[13] This is possible if $\lambda$ is too small to induce a spontaneous transition between two consecutive states $\nu\to\nu+1$. The transition takes place if, at the right time, the system receives an external signal $h_i=h_R\xi_i^{\nu+1}$ (recognition) or $h_i=h_c\eta_i$, where $\eta_i$ is uncorrelated with the stored patterns (counting).

One of us has proposed that the recognition of a sequence of external signals can be achieved, by the same mechanism, in a network in which each state is connected to several possible successors.[14] A simple realization of such a scheme is a network with two sequences of stored patterns $\{\xi_i^{1,\mu}\},\{\xi_i^{2,\mu}\}$. The transition term in the local field is

$$h_i^{tr}(t)=\lambda\sum(\xi_i^{1,\mu+1}+\xi_i^{2,\mu+1})[\tilde{S}_{1,\mu}(t-\tau)+\tilde{S}_{2,\mu}(t-\tau)]. \tag{6}$$

Thus, if the system is in state $(1,\nu)$, for example, it can go to either $(1,\nu+1)$ or $(2,\nu+1)$.

The cross terms in Eq. (6) allow for the two transitions. Which of them actually occurs will now be determined by the external signal, which is thereby recognized. It has been shown[14] by numerical simulations that there exists a range of parameters where such a network performs as a sequence recognizer, discriminating between the $2^p$ possible sequences of signals. The main point is that the external signal is too weak to induce the transitions by itself, so that recognition is a combined effect of previous learning and external inputs. Other schemes for recognition of temporal sequences have also been proposed.[8,15]

In the present paper we investigate the three modes of processing temporal sequences mentioned: retrieval, counting, and recognition.

The methods of Ref. 3 do not apply to nonsymmetric interactions like (2). Therefore, in order to get analytic results on the dynamics, we have studied a strongly diluted nonsymmetric version of the model. Such a version of the standard model ($\lambda=0$) has been introduced and solved recently.[16] This version does not necessarily represent a realistic network—it allows us to get an analytical understanding of the processes at work. It is known in general that diluted networks behave qualitatively in the same way as fully connected networks in the region of good retrieval. To check that this is also the case in the present application to sequences, we shall compare the results with numerical simulations on fully connected networks.

The interactions $J_{ij}$ are multiplied by random independent parameters $C_{ij}$ which can take the value 1 with probability $c/N$ and 0 with probability $1-c/N$, where $c$ is a finite number. Since $C_{ij}$ and $C_{ji}$ are independent variables, the interaction matrix is not symmetric. The retrieval of a pattern $\{\xi_i^\mu\}$ is represented by the time evolution of its thermal averaged overlap, $m_\mu(t)$, with the spin configuration. The time evolution is particularly simple in the limit $c\to\infty$ (keeping in mind that $N\to\infty$ first, and $c\ll\log N$). For parallel dynamics one has[16] in the "standard" case, $\lambda=0$,

$$m_\mu(t+1)=F_w(m_\mu(t)), \tag{7}$$

where

$$F_w(A)=\int\frac{dz}{(2\pi)^{1/2}}e^{-(z^2/2)}\tanh[\beta(A+\sqrt{w}z)], \tag{8}$$

and $w$ is the width of a "noise" due to random overlaps ($m_\nu\sim 1/\sqrt{N}$, $\nu\neq\mu$) with all the patterns which are not macroscopically polarized. A slightly different evolution equation exists for sequential dynamics, which, however, leads to the same fixed point.[16] In the standard case $w=p/c=\alpha$, and the network behaves as an associative memory for values of the parameters (small $T$ and small $\alpha$) such that (7) possesses a nontrivial fixed point.

We now proceed to discuss this same model in the case $\lambda\neq 0$.

*Retrieval.*—Let us first suppose that during the last $\tau$ time steps, the system was stabilized in the direction of one pattern $\mu=1$ (the self-consistency of this assumption will be checked later on): for $-\tau+1\le t\le 0$,

$$m_1(t)=Q^1;\quad m_\mu(t)=0,\quad \mu=2,\ldots,p. \tag{9}$$

Using the method of Ref. 16, one derives the recursion relation, for $1\le t\le\tau$,

$$m_\pm(t+1)=F_w(m_\pm(t)\pm\lambda Q^1), \tag{10}$$

where $m_\pm=m_1\pm m_2$ and $F_w$ is the sigmoidal function defined in (8). If, among the $p$ patterns, $q$ belong to sequences, the width of the noise is $w=\alpha[1+(q/p)\lambda^2]$. We shall discuss the case $q=p=\alpha c$, which in the thermodynamic limit corresponds to a finite number of infinitely long sequences. Results for other cases can be easily deduced from ours by the proper scaling of $\alpha$. The overlaps onto the other patterns $\mu=2,\ldots,p$ remain zero for $t\le\tau$. One must find the fixed points of (10) reached from the initial condition $m_\pm(0)=Q^1$. This gives the phase diagram shown in Fig. 1.

For large values of the transition parameter $\lambda$, *or large values of the noise* (by "noise" we mean both the thermal noise governed by the temperature and the internal noise, governed by $\alpha$, due to the existence of interference effects for an infinite number of patterns), the system makes a discontinuous transition towards a fixed point $m_1^*=0$, $m_2^*=Q^2$. This happens for $\lambda>\lambda_c(T,\alpha)$. The cross sections of this surface by planes of constant

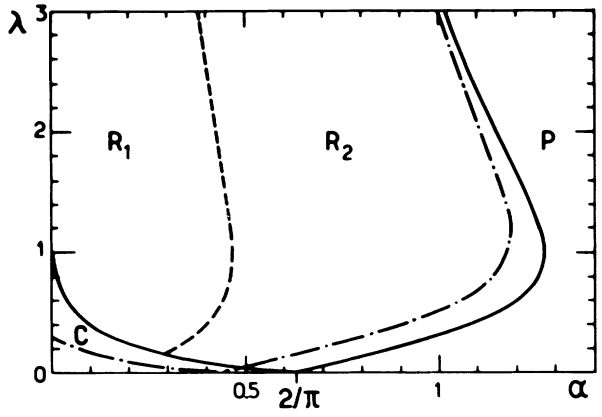


FIG. 1. Phase diagram in the $(\lambda,\alpha)$ plane for retrieval of temporal sequences. The constant-temperature curves, $T=0$ (full lines) and $T=0.5$ (dash-dotted lines) define three regions. In region $C$ the system remains in the previous state. This is the region of interest for the counting scheme. In region $P$ the system makes transitions but converges asymptotically to the paramagnetic state. Correct retrieval of the entire sequence is achieved in regions $R_1$ and $R_2$. The dashed curve defines a region $R_1$ in which the overlaps with the successive patterns exceed 0.95, at $T=0$. The $C$-$R$ boundary has been computed for an initial overlap $Q^1=0.9$.

temperature are shown in Fig. 1. For any value of $\lambda$, the transition from pattern 1 to pattern 2 takes place if the noise is strong enough. If $\lambda < \lambda_c(T,\alpha)$ the fixed point is at $m_1^* > m_2^*$.

If $\tau$ is large enough it is reasonable to assume that the system stabilizes into the fixed point before the next transition takes place (the convergence to the fixed point is exponentially fast away from the transition). Therefore the next transition $\mu = 2 \rightarrow 3$ is described by the same equation (10), where now $m_\pm = m_2 \pm m_3$, and the initial condition is $m_+ = m_- = Q^2$.

So, superimposed on the dynamics on short time scales (10) which describes the transition from one pattern to the next, there is a behavior on large time scales which gives the values of the overlaps $Q^\mu$ on the successive patterns. This long-time dynamics is a mapping $Q^\mu \rightarrow Q^{\mu+1}$ obtained by solution of the fixed-point condition

$$Q^{\mu+1} = F_w(Q^{\mu+1} + \lambda Q^\mu), \quad Q^{\mu+1} > 0. \tag{11}$$

The overlap $Q^\mu$ converges asymptotically towards a nonzero value $Q^*$ if and only if $(1+\lambda)F_w'(0) \geq 1$. This condition is realized if the noise is not too strong as can be seen in Fig. 1. It is possible to store more patterns in a sequence than separately, as in the case of $\lambda = 0$. This is possible since a nonzero $\lambda$ increases not only the noise, but also the signal which stabilizes the patterns. However, the transition to $Q^* = 0$ is continuous and one should keep away from the transition line in order to have a sizable $Q^*$.

The transition from the region $C$ to $R$ in Fig. 1 compares qualitatively with simulations on fully connected networks. The main difference between strongly diluted and fully connected networks is the behavior when the noise increases. In the first case sequences are retrieved with an averaged overlap on successive patterns which decreases gradually to zero at the right-hand curves in Fig. 1. In the latter case this overlap drops sharply to zero from $m \simeq 1$ at a critical $\alpha_c(\lambda)$. Results of simulations on fully connected networks of $N = 200$, at $T = 0$, indicate that $\alpha_c = 0.22$–$0.25$ for $\lambda = 1$ and $\alpha_c = 0.26$–$0.3$ for $\lambda = 2$–$3$. More extensive simulations are needed to derive the full curve more accurately.

*Counting.*—This system can be used to count chimes as suggested in Ref. 13. Suppose the system is started as before, polarized in the direction of pattern 1 [see (9)]. But now the parameters are such that $\lambda < \lambda_c(T,\alpha)$ (region $C$ in Fig. 1), so that the transition to pattern 2 does not take place. The additional noise due to a chime arriving at time $t_1$ can trigger the transition. This will add to the local field of (4) a random term $h_c \eta_i \delta_{t,t_1}$ ($\eta_i = \pm 1$), uncorrelated with any of the patterns, and $h_c$ is the amplitude of the chime. At time $t_1$, (10) is modified to

$$m_\pm(t_1+1) = \tfrac{1}{2} \sum_{\eta = \pm 1} F_w(m_\pm(t_1) \pm \lambda Q^1 + \eta h_c). \tag{12}$$

The effect of $h_c$ can drive the transition to pattern 2 (which is the internal representation of the cardinal following the one represented by pattern 1) if $h_c > h_c^*(\lambda, T, \alpha)$. If noise is initially present ($T$ and/or $\alpha \neq 0$), a small signal (small $h_c$) is enough to induce the transition and allow for the chime to be counted, while for $T = \alpha = 0$ one needs[13] $h_c \geq 1 - \lambda$.

This counting scheme does not require the chimes to be periodic. There is a lower bound $\hat{t}_L$ on the interval, $t_k - t_{k-1}$, between two chimes: $\hat{t}_L$ must be large enough so that the system is ready to jump to state $k+1$ at the moment $t_k$. If $t_{\rm th}$ is the thermalization time, one needs $\hat{t}_L \geq \tau + t_{\rm th}$. The smaller $\lambda$, the slower the thermalization, the larger $t_L$. On the other hand, there is no upper bound on $t_k - t_{k-1}$.

*Recognition.*—The recognition of one sequence can be achieved by the same mechanism as counting, if the system receives at time $t_k$ a signal $h_R \xi_i^{\mu = k+1} \delta_{t,t_R}$ which tends to drive the transition to pattern $k+1$. Recognition takes place if $h_R > h_R^*(\lambda, T, \alpha)$. We have found that $h_c^*(\lambda, T, \alpha) \geq h_R^*(\lambda, T, \alpha)$ [a typical example is $h_c^*(0.3, 0, 0.1) = 0.64$ and $h_R^*(0.3, 0, 0.1) = 0.42$], so that there is a range of parameters in which the system discriminates between signals corresponding to random and stored patterns. The exception is $T = \alpha = 0$, for which $h_c^* = h_R^* = 1 - \lambda$.[13]

We now turn to the model of recognition in which the system can bifurcate between two sequences, as defined in (6). We start again with a state polarized on a single pattern: $m_{11} = Q$, $m_{21} = 0$, $m_{i\mu} = 0$ ($i = 1,2$; $\mu = 2, \ldots, p$). An external signal of strength $h$, conjugate to one of the next patterns in the two sequences, (1,2) or (2,2), arrives at time $t_1$. To study the dynamics of the transition, in the diluted model, we have to iterate the three coupled equations for the evolution of the thermally averaged

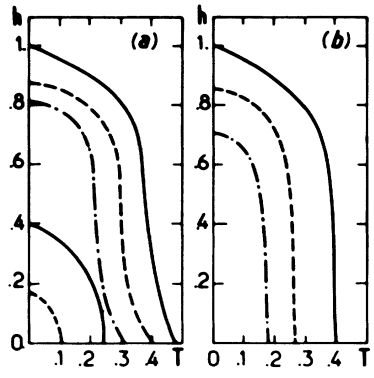


FIG. 2. Phase diagram describing a single transition in the recognition model for (a) $\lambda = 0.3$ and (b) $\lambda = 0.5$, and for $\alpha = 0$ (full lines), $\alpha = 0.03$ (dashed lines), and $\alpha = 0.05$ (dash-dotted lines). The lower curves in (a) define regions (to the left of these curves) in which the system remains in the initial state. All other curves separate between a region of transitions to a mixture of three states (on the left side) and a region of correct recognition.

overlaps $m_{11}(t)$, $m_{12}(t)$, and $m_{22}(t)$. The noise parameter is $w=\alpha(1+2\lambda^2)$, where, now $\alpha=2p/c$. Without writing down the equations, we shall discuss their stationary solutions.

At very low values of $T$ and $\alpha$ there is a region in which the network remains in the initial state [the two inner curves in Fig. 2(a)]. At $T=0$ and $\alpha=0$, this occurs for $h<1-2\lambda Q$. For $1-2\lambda Q<h<1$, the fixed point represents an equal mixture of the three patterns. When either $T$ or $\alpha$ is increased, the overlaps with the three patterns become gradually unequal, $m_{12}>m_{22}>m_{11}$ [when the external signal is conjugate to pattern (1,2)], but all of them are macroscopic. At a critical value of $T$ (or $\alpha$) there is a sharp transition to a state with $m_{11}=0$, $m_{12}\simeq 1$, $m_{22}\simeq 0$. This is the region where the network recognizes an external signal. Several examples of combinations of parameters where this occurs are shown in Fig. 2. The overall behavior is, qualitatively and even semiquantitatively, reproduced in numerical simulations on fully connected networks. Again noise is very important: At zero temperature the system can easily get stuck into a mixture of three patterns ($m_{11}\sim m_{21}\sim m_{22}\sim 0.5$), while the adjunction of thermal noise helps it to make the proper recognition.

Let us make a brief remark on the long-time behavior. When the system is in the region where a signal is well recognized, and $T$ (or $\alpha$) is increased further, then $m_{12}$ decreases and $m_{22}$ increases until they continuously become equal. The rate at which this happens (which depends strongly on $\lambda$, through its effect on $w$, and weakly on $h$) will determine the level of noise ($T$ or $\alpha$) at which the system can recognize a long sequence and not just a single signal. The discussion of this point and comparison with numerical simulations on fully connected systems will be presented elsewhere. Let us only mention, to give a general feeling, that in simulations for $N=200$, $\rho=10$ ($\alpha=0.1$), $\lambda=0.3$, $h=0.8$, and $T=0$ the system makes a correct recognition of the entire sequence, with averaged $m>0.95$, in 85% of the cases.

Another point to be considered is the dependence on the time constants involved—the delay time $\tau$ and the interval between two consecutive signals. The arguments, presented in the section on counting, for the lower bound on the interval between two signals apply here as well. In the region of parameters where the noise can induce the transition, there is also an upper bound. The external signal has to arrive before the combined action of $\lambda$ and noise drives the system out of the stabilized pattern to either one of the possible successors, selected by fluctuations in the overlaps, or to a mixture of the two.

The main point of this paper was to show that the model of diluted neural networks, proposed in Ref. 16, can be extended to get an analytic understanding of networks with temporal sequences of patterns. Furthermore, the results shed light on the corresponding behavior in fully connected networks. An interesting result concerns the role of noise, both thermal and internal. Noise helps to retrieve temporal sequences stored with lower values of $\lambda$, and count or recognize weaker signals. In the context of retrieval, such a role of noise was already pointed out in Ref. 6.

We are grateful to D. Amit, J. P. Changeux, J. P. Nadal, and G. Toulouse for discussions on temporal sequences in neural networks. One of us (H.G.) is grateful to the faculty and staff of the Laboratoire de Physique de l'Ecole Normale Supérieure, to the Ecole Supérieure de Physique et Chimie Industrielle, and to the University of Paris VI, for their hospitality. Laboratoire de Physique Théorique de l'Ecole Normale Supérieure Laboratoire Propre du Centre National de la Recherche Scientifique is a Laboratoire Associé à l'Ecole Normale Supérieure et à l'Université de Paris-Sud.

[a]Present address: The Racah Institute of Physics, The Hebrew University, Jerusalem, Israel.

[b]Postal address: 24 rue Lhomond, 75231 Paris, Cedex 05, France.

[1]J. J. Hopfield, Proc. Natl. Acad. Sci. USA **79**, 2554 (1982), and **81**, 3088 (1984).

[2]W. A. Little, Math. Biosci. **19**, 101 (1974); W. A. Little and G. L. Shaw, Behav. Biol. **14**, 115 (1975).

[3]D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985), and Ann. Phys. (N.Y.) **173**, 30 (1987).

[4]H. Sompolinsky, Phys. Rev. A **34**, 2571 (1986).

[5]P. Peretto and J. J. Niez, in *Disordered Systems and Biological Organizations*, edited by E. Bienenstock, F. Fogelman-Soulié, and G. Weisbuch (Springer-Verlag, Berlin, 1986), p. 171.

[6]H. Sompolinsky and I. Kanter, Phys. Rev. Lett. **57**, 2861 (1986).

[7]D. Kleinfeld, Proc. Natl. Acad. Sci. USA **83**, 9469 (1986).

[8]S. Dehaene, J.-P. Changeux, and J.-P. Nadal, Proc. Natl. Acad. Sci. USA **84**, 2727 (1987).

[9]J. Buhmann and K. Schulten, Europhys. Lett. (to be published).

[10]S. Grossberg, Stud. Appl. Math. **44**, 135 (1970), and J. Cybern. **1**, 28 (1971).

[11]T. Kohonen, *Self-Organization and Associative Memory* (Springer-Verlag, Berlin, 1984).

[12]P. Kanerva, "Self-Propagating Search: A Unified Theory of Memory" (to be published).

[13]D. J. Amit, to be published.

[14]H. Gutfreund, to be published.

[15]D. W. Tank and J. J. Hopfield, Proc. Natl. Acad. Sci. USA **84**, 1896 (1987).

[16]B. Derrida, E. Gardner, and A. Zippelins, Europhys. Lett. **4**, 167 (1987).