

Disambiguation of Pattern Sequences with Recurrent Networks

Ali A. Minai

Dept. of Electrical and Computer Engg.
University of Cincinnati
Cincinnati, OH 45221-0030

Geoffrey L. Barrows

Department of Electrical Engineering
Stanford University
Stanford, CA 94305

William B. Levy

Department of Neurosurgery
University of Virginia
Charlottesville, VA 22908

Abstract

Recently, there has been great interest in using neural networks to learn sequences of patterns. This is obviously very important from a cognitive point of view. In this paper, we show how a simple network with a hippocampal-like structure can be used to learn complex stimulus sequences. Specifically, we train the system, which has a one-step recurrent dynamics, to disambiguate sequences with temporal overlaps of more than one step. Usually, this is done either through delay lines or by means of capacitive effects. We show that a significant population of unforced, recurrently activated neurons in the system can enable the system to disambiguate quite well over several time steps.

1. Introduction

Recurrently connected networks of neuron-like elements have been widely used for associative storage of pattern sequences (Amari, 1972; Fukushima, 1973; Sompolinsky & Kanter, 1986; Kleinfeld, 1986; Buhmann & Schulten, 1987; Coolen & Gielen, 1988; Bauer & Krey, 1990; Jordan, 1986; Mozer, 1989; Elman, 1990; Reiss & Taylor, 1991; Heskes & Gielen, 1992; Bartholomeus & Coolen, 1992). The ability to learn sequences is obviously important in the cognitive context where the brain must constantly process time-varying information. In this paper, we consider the problem of disambiguating stimulus sequences with temporal overlap. Learning such sequences requires that context information be remembered over an extended period. We show that a simple network, inspired by the structure of the mammalian hippocampus, can learn sequences with long temporal overlap without any explicit memory mechanism beyond one-step recurrence.

We consider sequences with one pattern per step and use a sparsely connected recurrent network with modified hebbian learning. However, we *do not* force all neurons externally, allowing a portion of them to fire freely depending on the recurrent activation. As we discuss below, this is a key feature of the system

2. First-Order and Higher-Order Sequences

An important issue for sequence learning is that of order: how many previous states does a transition depend on? In the simplest case, each pattern has a unique successor, giving a first-order finite-state machine. A more interesting situation arises when the same pattern (or sub-sequence) can have different successors depending on several previous states. This creates a higher-order finite-state machine and leads to the problem of *disambiguation*: learning the correct transition when it depends on more than just the current state (Fukushima, 1973; Reiss & Taylor, 1991).

Problems involving higher-order situations have typically been solved using two devices: an explicit multiplicity of delays (Fukushima, 1973; Coolen & Gielen, 1988; Heskes & Gielen, 1992; Bartholomeus & Coolen, 1992); or capacitive effects (Reiss & Taylor, 1991). In both situations, the aim is to provide information from the past in order to facilitate the decision at the current transition point. We show that the same effect can be obtained by allowing a small amount of unforced recurrent activity in the network, provided that a sizeable proportion of neurons is never directly activated by any pattern in the sequence.

3. The Biological Motivation

Our study of sequence learning in neural systems was motivated largely by a desire to understand the mammalian hippocampus. We have previously suggested (Levy, 1989; Minai & Levy, 1993b) that the hippocampus might be involved in the storage, recall, and prediction of sensory sequences. In the context of this hippocampal hypothesis, we constructed a three layer system: an input layer, corresponding roughly to the entorhinal cortex and the dentate gyrus; a recurrent associative layer, corresponding to the CA3 region; and an output layer analogous to the CA1. The input layer provided strong reinforcement to the CA1 and stimulated *some of the neurons* in CA3. It also sent non-specific inhibitory signals to both layers to act as a normalization mechanism. The CA3 layer sent specific,

randomly placed excitatory and non-specific inhibitory connections to itself and to the CA1 layer (see Figure 1).

The input layer provided stimulus sequences to the CA3-CA1 system. The CA3-CA3 and the CA3-CA1 excitatory weights were modified via an associative rule to learn the sequences. The system's goal was to complete a learned sequence given an initial fragment. Sequence learning occurred primarily in the CA3 layer, which constructed an encoding of the dynamics implicit in the stimulus. The CA3 neurons not directly stimulated by the input were left free to encode contextual representations. The CA1 layer acted essentially as a pattern recognition/categorization layer, recovering the original sequence patterns from the CA3 recodings.

4. Network Specification and Problem Definition

We used a network model which builds on one that we have previously investigated (Minai & Levy, 1993 a,b,c). A network had N inputs and outputs and consisted of three layers of binary neurons: an input layer, I ; a recurrent CA3 layer; and a CA1 output layer. The input layer consisted of N dummy units which simply distributed the N network inputs to the other two layers. Thus, the output, x_i , of the i th input layer neuron was equal to the i th network input. The recurrent layer consisted of $n \geq N$ binary (0/1) primary neurons with identical firing thresholds, θ . The neurons were interconnected via a Bernoulli process: each neuron i had probability p of receiving a modifiable excitatory connection from each neuron j in the recurrent layer. The presence of such a connection was indicated by the binary variable c_{ij} . Neurons 1 through N also received strong one-to-one excitatory inputs, x_i , from the input layer through a synapse of fixed strength v . Inhibition was mediated by a single interneuron that received input from all primary neurons in the layer and all inputs, x_i , and then provided an identical shunting conductance proportional to its input to all primary neurons. At time t , taking $w_{ij}(t)$ as the excitatory weight from neuron i to j , K_I as the fixed inhibitory weight from the input layer, K_R as the fixed weight for feedback inhibition, $m(t)$ as the number of active neurons in the recurrent layer, and $s(t)$ as the number of active inputs, the excitation y_i of neuron i was given by:

$$y_i(t) = Input_i(t) + \sum_{j \in R} c_{ij} w_{ij} z_j(t-1) - [K_I s(t) + K_R m(t-1)] \quad (1)$$

where $Input_i(t) = vx_i(t)$ if $1 \leq i \leq N$ and 0 otherwise. We defined $y_i(t) = 0$ for all i if $s(t) = m(t-1) = 0$. The output of the neuron was calculated as:

$$z_i(t) = \begin{cases} 1 & \text{if } y_i(t) \geq \theta \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

It was assumed that input weight v was strong enough that $x_i(t) = 1 \Rightarrow z_i(t) = 1$. However, neurons with no external input were *not* forced to the zero state and could be fired through feedback connections.

The CA1-like output layer consisted of N primary neurons which received random modifiable connections, a_{ij} , from the recurrent layer and strong, non-modifiable, one-to-one connections of magnitude b from the input layer. They also got feed-forward inhibition via an interneuron that received input from the recurrent and input layers. The equations for neurons in the output layer were:

$$y_i(t) = bx_i(t) + \sum_{j \in R} c_{ij} a_{ij} z_j(t) - [C_I s(t) + C_R m(t)] \quad (3)$$

$$z_i(t) = \begin{cases} 1 & \text{if } y_i(t) \geq \phi \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where C_I and C_R were inhibitory weights from the input and CA3 layers, respectively, and ϕ was the firing threshold. The CA3-CA1 weights began at zero so that, initially, the learning from the recurrent to the output layer was driven by the network input. The purpose of the output layer was to recover the original patterns from the context-dependent representations developed by the recurrent layer, effectively implementing an inverse mapping.

5. The Problem

The task investigated was to learn to disambiguate two higher-order sequences of n -dimensional binary patterns. Temporal overlap was created between the sequences by including identical sub-sequences in both. The length of this overlap was varied from 0 to 4 to evaluate disambiguation performance. The sequences were constructed from a

repertoire of 20 mutually orthogonal patterns, labeled A through T, as follows:

0 step overlap: ABCDEFGH and IJKLMNOP

1 step overlap: ABQDEFGH and IJQLMNOP

2 step overlap: ABQREFGH and IJQRMNOP

3 step overlap: ABQRSFGH and IJQRSNOP

4 step overlap: ABQRSTGH and IJRSTOP

At least the first two and the last two patterns in each sequence pair were distinct and mutually orthogonal. The network was evaluated by its ability to reproduce the final pattern of a sequence at step 8 given the initial pattern at step 1.

5.1. The Basic Premise

The basic premise of this investigation was that simple one-step associative learning could imprint higher-order sequences in recurrent networks if some neurons not directly stimulated by patterns were also allowed to fire. This *auxiliary activity* creates distinguishable, context-dependent secondary representations of the stimulus patterns and allows the correct transitions to be learned. In essence, auxiliary activity plays the same role as hidden neurons do in a supervised learning situation. It allows the received dynamics of the input stimulus to be mapped to a higher-dimensional space where degeneracies can be distinguished in the extra dimensions. Auxiliary activity can come from noise added to patterns in off-line training, or be produced through excitatory feedback if the training is on-line. The very long transients produced through recurrence (Minai & Levy, 1993 a,b,c) can be used to this end.

5.2. The Training Process

For training, the network was repeatedly stimulated by the two sequences and the CA3-CA3 and CA3-CA1 synapses were allowed to modify. We tried two different learning rules in the CA3 layer:

Postsynaptic Rule: $w_{ij}(t) = w_{ij}(t-1) + \epsilon z_i(t) [z_j(t-1) - w_{ij}(t-1)]$

Presynaptic Rule: $w_{ij}(t) = w_{ij}(t-1) + \epsilon z_j(t-1) [z_i(t) - w_{ij}(t-1)]$

where ϵ was a small learning rate parameter. Only the presynaptic rule was used for the CA1 layer. The relative pros and cons of the rules are discussed in Sections 5.4 and 6. Both rules are biologically plausible (see Levy (1982); Levy, Colbert & Desmond (1990) for the postsynaptic rule and Fujii et al (1991); Dudek & Bear (1992); Mulkey and Malenka (1992) for the presynaptic.) At the beginning of each sequence presentation, the state of the network was reset to 0, so that the sequences did not merge together.

5.3. Performance Evaluation

To test recall, all neurons were reset to 0 and the system was then stimulated with the first pattern of a learned sequence. The network was then allowed to relax without further input for 7 time steps; its output over the final step of the sequence (step 8) was compared with the corresponding true pattern in the sequence to evaluate performance on that sequence. The process was then repeated with the other sequence. The completion performance on a sequence was judged by calculating two measures:

- 1) The number of correct neurons on
- 2) The number of incorrect neurons on.

Each measure was averaged over both sequences and over ten networks, each with different, randomly generated CA3-CA3 and CA3-CA1 connection matrices. It should be noted that the patterns used were very sparse (10 active bits in each), so the correct number of active neurons at any step was far lower than the number that should have remained inactive. Thus, in a 10-out-of-100 pattern, if 9 correct neurons were active and 9 incorrect ones, the *proportion* of correct to incorrect firing would be 9 : 1, not 1 : 1. We chose to plot absolute numbers rather than proportionate values because they represent a more exacting criterion.

5.4. Choice of Learning Rules

Two different situations were investigated in our simulations. In the first set of experiments, the CA3-CA3 weights were modified using the pre-synaptic rule, while in the second series, the post-synaptic rule was used. The pre-synaptic rule was used in both cases for the CA3-CA1 weights. The reason can be appreciated by considering the

difference between the two rules, which is basically in their depressive aspects. Suppose two pre-synaptic patterns, V and W, map into the same post-synaptic pattern, X. If a post-synaptic rule is used, it sets up *competition* between the synapses used by V and those used by W: when the V-X pair is active, the W-X synapses are depressed, and vice-versa when W-X is active. The result is that neither V nor W gets properly associated with X, whereas both should have. With the pre-synaptic rule, the competition is not between candidate pre-synaptic patterns but between prospective post-synaptic ones, i.e., if X elicits Y or Z in different situations, the X-Y and X-Z synapses compete. For synapses from CA3 to CA1, the latter situation should not arise, since the CA3 should not form identical representations for two distinct input/output patterns. The former situation, however, occurs routinely, since two convergent sub-sequences (e.g., ABQ and IJQ) will usually lead to different CA3 patterns but should both elicit Q in CA1. It is, therefore, clear that the pre-synaptic rule is the rule of choice for CA3-CA1 synapses. However, this is not so clear for the CA3-CA3 synapses where the association is temporal, and it is possible to have convergent (ABQ and IJQ) and divergent (TGH and TOP) situations.

6. Simulation Results and Discussion

The results from the simulations are shown in Figure 2 (a & b). In each case, the CA3 layer was 200 neurons wide and each neuron received excitatory connections from 40 other randomly chosen neurons. The input bus was also 200 neurons wide, but its effective size depended upon the number of different patterns in the two sequences. At most 160 neurons were used (in the 0-overlap situation) and at least 120 (in the 4-overlap case). The CA1 mirrored the input layer, with 200 neurons and a variable effective size. The input layer projected in a 1-1 manner to the CA3 and the CA1 with powerful "sure-fire" synapses. The CA3 projected to CA1 randomly with a 40% chance of connection between two specific neurons. The CA3-CA3 weights were initially set to values that allowed some recurrently generated CA3 activity. However, the CA3-CA1 weights were initially zero so that the CA1 learned only through reinforcing input layer activity. Each data point was averaged over both sequences in the corresponding pair and over ten random networks.

The results show several interesting features. The most notable is that our system, without any explicit long delay memory mechanism, is able to disambiguate quite well across several steps of overlap. This is especially true with the pre-synaptic rule. In this case, even an overlap of 4 activates about 45% of the correct neurons and only 2% of the incorrect ones.

Another notable feature is the manner in which the two learning rules affect the error with increasing temporal overlap. The pre-synaptic rule degrades in terms of correct neurons on but holds the number of spurious firings down. The post-synaptic rule does the opposite. This can be explained by considering the point of decision in the network's task. Since performance is judged by the ability to reach the correct one of two possible terminations, the critical point is when the system emerges from the temporal overlap and opts for one terminal direction or the other. For example, in the ABQREFGH/IJQRMNOP situation (overlap = 2), the critical decision is made at the R→E/R→M transition. This is also where the effect of temporal overlap is maximal and the CA3 representations of the two contexts most alike. Let X be the final pattern in a temporal overlap of length q , to be followed by Y or Z. Since X can be arrived at in two contexts, it will have two CA3 representations, one of which should elicit Y and the other Z. Call these representations X_1 and X_2 , respectively, and let X' be their intersection. Then the size of X' is an increasing function of q . Now, the post-synaptic rule, at this point, associates X_1 with Y and X_2 with Z. More importantly, the active bits in X' are associated *equally strongly* with *both* Y and Z, without competition. As the temporal overlap becomes longer, X' approaches both X_1 and X_2 . Thus, there will be an increased tendency to elicit *both* Y and Z regardless of context, leading to the situation in Figure 2(b). Note also that this problem cannot be fixed simply by using a competitive firing mechanism like firing the 10 most excited neurons.

In the presynaptic case, however, there is competition at the point of decision, so that synapses from X' to Y and Z are kept relatively weak, while synapses from $X_1 - X_2$ to Y and from $X_2 - X_1$ to Z are strengthened. Thus, the decision is made on the basis of the *differentiated* parts of the two CA3 representations rather than on the entire representations as in the post-synaptic case. In this sense, the pre-synaptic rule is the logically correct choice for learning disambiguation. However, since the differentiated parts of the CA3 representations become smaller with increasing q , the total activity in the network also declines. This effect is apparent in Figure 2(a). However, this situation can be improved by a competitive firing mechanism.

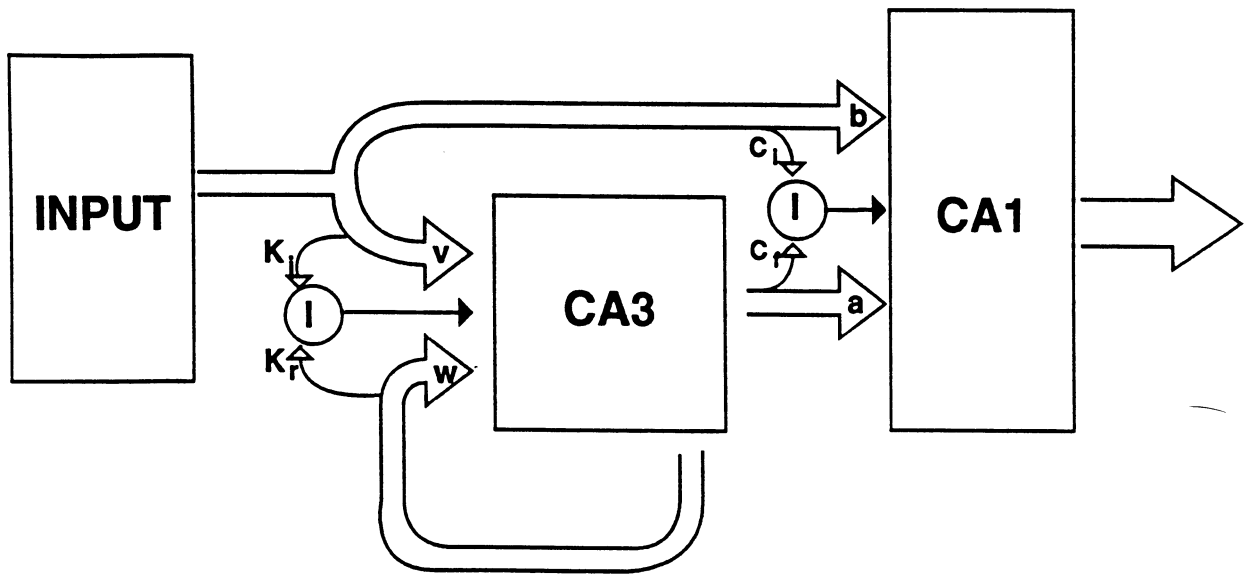


Figure 1: Schematic of the system components and connections. The wide arrows indicate specific connections, and the thin line arrows indicate non-specific inhibitory pathways feeding into and out of interneurons (I). Open arrowheads are excitatory synapses, and solid ones are inhibitory synapses. Only connections w and a are modifiable.

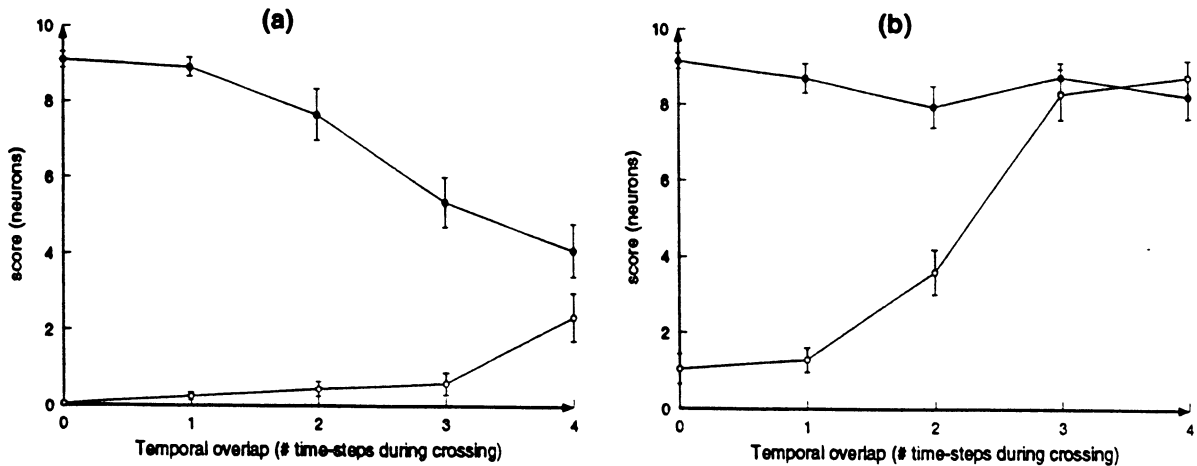


Figure 2: The performance of the system on sequences with different temporal overlaps. The filled circles indicate the number of correctly firing neurons and the open circles show the number of spurious firings. Each data point is averaged over ten randomly generated networks and over the sequence pair. Error bars show one standard deviation. Every network was trained with 200 passes over both sequences. Graph (a) depicts the case where CA3 weights were modified by the pre-synaptic rule. Good parameter settings were found by trying different combinations on the 3-step overlap. These settings were then used for all cases. CA3 weights were initially set to 0.36. The learning rate for all weights was 0.02. Graph (b) shows the results for the case where CA3 weights were trained with the post-synaptic rule. Attempts to find good parameter settings with 3-step overlap case failed in this case because the network did very poorly on this problem. Thus, the 2-step overlap case was used to find good parameter values. The initial CA3 weight was 0.48, and the learning rate was 0.02.

Acknowledgements: This research was supported in part by the following grants to WBL: NIMH MH00622; NIMH MH48161; NSF MSS-9216372; and EPRI RP8030-08. It was also supported by the Department of Neurosurgery, University of Virginia, Dr. John A. Jane, Chairman. The authors would like to thank Dawn Adelsberger-Mangan for her constructive comments and Colin Prepscius for his help with the manuscript.

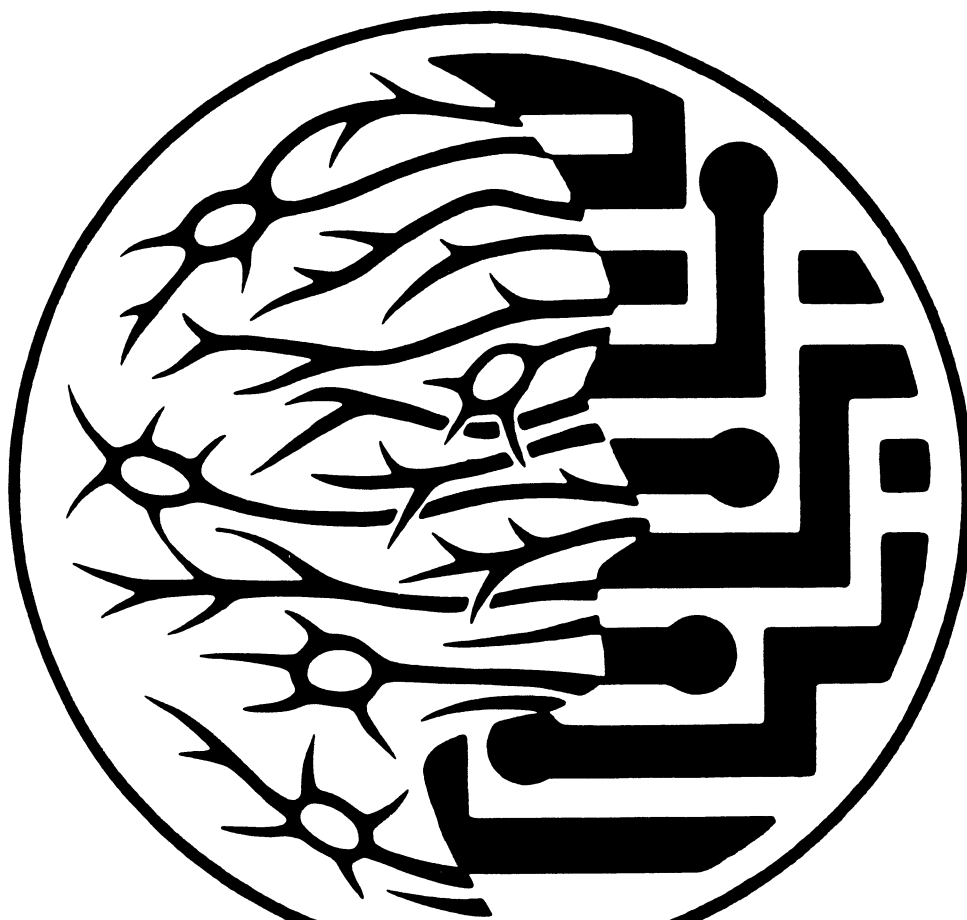
- S. Amari (1972) "Learning Patterns and Pattern Sequences by Self-Organizing Nets of Threshold Elements", *IEEE Trans. on Computers*, C-21, 1197-1206.
- M. Bartholomeus & A.C.C. Coolen (1992) "Sequences of Smoothly Correlated Patterns in Neural Networks with Random Transmission Delays", *Biological Cybernetics*, 67, 285-290.
- K. Bauer & U. Krey (1990) "On Learning and Recognition of Temporal Sequences of Correlated Patterns", *Z. Phys. B - Condensed Matter*, 79, 461-475.
- J. Buhmann & K. Schulten (1987) "Noise-Driven Temporal Association in Neural Networks" *Europhys. Lett.*, 4, 1205-1209.
- A.C.C. Coolen & C.C.A.M. Gielen (1988) "Delays in Neural Networks", *Europhys. Lett.*, 7, 281-285.
- S.M. Dudek & M.F. Bear (1992) "Homosynaptic Long-Term Depression in Area CA1 of Hippocampus and Effects of N-Methyl-D-Aspartate Receptor Blockade", *Proc. Nat. Acad. Sci. USA*, 89, 4363-4367.
- J.L. Elman (1990) "Finding Structure in Time", *Cognitive Science*, 14, 179-211.
- S. Fujii, K. Saito, H. Miyakawa, K. Ito & H. Kato (1991) "Reversal of Long-Term Potentiation (Depotentiation) Induced by Tetanus Stimulation of the Input to CA1 Neurons of Guinea Pig Hippocampal Slices", *Brain Res.* 555, 112-122.
- K. Fukushima (1973) "A Model of Associative Memory in the Brain" *Kybernetik*, 12, 58-63.
- T.M. Heskes & S. Gielen (1992) "Retrieval of Pattern Sequences at Variable Speeds in a Neural Network with Delays", *Neural Networks*, 5, 145-152.
- M.I. Jordan "Attractor Dynamics and Parallelism in a Connectionist Sequential Machine", *Proc. 8th Ann. Conf. of the Cog. Sci. Soc.*, 531-546.
- D. Kleinfeld (1986) "Sequential State Generation by Model Neural Networks", *Proc. Natl. Acad. Sci. USA*, 83, 9469-9473.
- W.B. Levy (1982) "Associative Encoding at Synapses", *Proc. Fourth Ann. Conf. Cognitive Sci. Soc.*, 135-136.
- W.B. Levy, C.M. Colbert & N.L. Desmond (1990) "Elemental Adaptive Processes of Neurons and Synapses: A Statistical/Computational Perspective", in: *Neuroscience and Connectionist Theory*, M. Gluck & D. Rumelhart (eds.), Hillsdale, NJ, Lawrence Erlbaum Associates, 187-235.
- A.A. Minai & W.B. Levy (1993a) "Predicting Complex Behavior in Sparse Asymmetric Networks", in: *Advances in Neural Information Processing Systems 5*, S.J. Hanson, J.D. Cowan & C.L. Giles (eds.), San Mateo, CA, Morgan Kaufmann, 556-563.
- A.A. Minai & W.B. Levy (1993b) "The Dynamics of Sparse Random Networks", *in press*.
- A.A. Minai & W.B. Levy (1993c) "Setting the Activity Level in Sparse Random Networks", *in press*.
- M.C. Mozer (1989) "A Focused Backpropagation Algorithm for Temporal Pattern Recognition", *Complex Systems*, 3, 349-381.
- Mulkey, R.M. & Malenka, R.C. (1992). Mechanisms underlying induction of homosynaptic long-term depression in area CA1 of the hippocampus, *Neuron*, 9: 967-975.
- M. Reiss & J.G. Taylor (1991) "Storing Temporal Sequences", *Neural Networks*, 4, 773-787.
- H. Sompolinsky & I. Kanter (1986) "Temporal Association in Asymmetric Neural Networks", *Physical Rev. Lett.*, 57, 2861-2864.

WORLD
CONGRESS
ON NEURAL
NETWORKS-
SAN DIEGO

1994 INTERNATIONAL
NEURAL NETWORK SOCIETY
ANNUAL MEETING

VOLUME 4

TOWN & COUNTRY HOTEL
SAN DIEGO, CALIFORNIA USA
JUNE 5-9, 1994



Copyright © 1994 by Lawrence Erlbaum Associates, Inc., and INNS Press, held jointly. All rights reserved. No part of this book may be reproduced in any form, by photostat, microform, retrieval system, or any other means, without the prior written permission of the publisher.

Lawrence Erlbaum Associates, Inc., Publishers
365 Broadway
Hillsdale, New Jersey 07642

ISBN 0-8058-1745-X

Books published by Lawrence Erlbaum Associates are printed on acid-free paper, and their bindings are chosen for strength and durability.

Printed in the United States of America