Latent Attractor Selection in the Presence of Irrelevant Stimuli

Simona Doboli
Computer Science Department
Hofstra University
Hempstead, NY 11549

Ali A. Minai
Complex Adaptive Systems Laboratory
ECECS Department
University of Cincinnati
Cincinnati, OH 45221

Abstract

Latent attractor networks are recurrent neural networks with weak attractors that bias the network's response to external stimuli but never fully manifest themselves. Such networks have been used to model context-dependent place representations in the hippocampus [5], and to encode context-dependent stimuli in neural networks [3]. In the original latent attractor model, each attractor was triggered by a unique *context pattern* **representing a stimulus that uniquely identified the context of the subsequent episode. This model was later extended to the case where contexts were triggered progressively by the sequential presentation of several stimulus patterns without regard to order. In this paper, we describe a network model that can select contexts even if the triggering stimulus patterns are interspersed among patterns irrelevant to context selection. This is closer to the way such a process would occur cognitively, where contexts are typically recognized based on a subset of sequentially perceived identifiers or cues among a larger set of perceived items.**

1 Introduction

In most cognitively interesting situations, the meaning of stimuli depends on context, which is typically specified at some earlier time. Two types of contexts can be distinuished depending on how far in the past the context information is given *relative to the current time*. The first one — which we call *Type I Context* — occurs when the system's response (output) at time t depends on stimuli (inputs) presented in an immediately preceding time window. Examples of such context appear in dynamical systems and in applications such as speech processing and word recognition. It is embodied in autoregressive models or finite-state machines and can be learned by recurrent neural networks where past states are fed back to the network [11, 13, 12].

The second type of context, which we term *Type II* or *episodic context*, arises when the information about context is given transiently at *a particular time* in the past. Examples of such context are found in social situations, in recognition of spatial environments, task specifications, etc., where the context is typically specified at the beginning of the episode and continues to be in force for its entire duration — even if the episode lasts a long time. As a concrete example, consider a situation when a set of identical rooms (e.g., in a hotel) are distinguished only by the number on the door. Looking at this number at the time of entry sets the context for the interpretation of subsequent experiences in the room. Episodic context dependence is more difficult to address using simple decaying feedback because the information on context is specified at a fixed time which grows increasingly remote from the present. This requires the system to "latch" information which is difficult even in recurrent networks [1, 12].

Experiments suggest that the hippocampal region of the brain in rodents constructs distinct representations of similar or even identical environments depending on episodic context. We have proposed a class of networks called *latent attractor networks* as a paradigm to explain how episodic context-dependent representations can arise without resorting to off-line or external biasing [18, 5, 3, 6]. Latent attractor networks are recurrent neural networks with competitive firing that embed patterns of activity as attractors using associative learning. However, the recurrent connections are not strong enough to stabilize the activity patterns autonomously. In the original formulation, each attractor is associated with a specific external stimulus pattern called a *context pattern*. When the context pattern is presented to the network, it disproportionately activates neurons in the active set of the associated attractor. This sets up a stable bias onto this set via the recurrent connections so that subsequent external stimuli — not explicitly associated with any particular attrac-

tor — also produce response patterns whose activity lies mainly in the active set of the chosen attractor. Thus, the system's responses to stimuli are conditioned by the original context pattern long after the context pattern itself is gone. This situation lasts until an external stimulus associated with another context/attractor is presented to the network [5].

In the paradigm described above, the context patterns represent the stimuli that set the context for an episode (e.g., looking at the number on a door). However, in typical situations, context is not set by a single stimulus, but by a conjunction of stimuli (e.g., objects in a room). For each episode with the same context, these stimuli may appear in different order and may be interspersed with other irrelevant stimuli. Furthermore, any one stimulus may be part of the combination indicating several contexts; it is the combination which denotes the specific context. Thus, as a cognitive system scans over the context-setting stimuli, a unique context would only emerge gradually rather than instantaneously. Indeed, until the context is uniquely identified, the system's response may be compatible with several choices.

In a previous study, we have considered the gradual activation of latent attractors by the presentation of a stimulus sequence [8]. In this paper, we extend this to include the presence of random, irrelevant stimuli in the context sequence. The problem of encoding episodes — in this case contexts — by a set of inputs presented sequentially to a neural system is relevant in other situations as well. For example in the case of storing information hierarchically in a neural network, several pieces of information at a lower level define a single entity at a higher level. Whenever the higher concept — for example, the title of a movie — cannot be recalled immediately, the sequential presentation, or recall of lower level concepts — like topic, actors, type, etc. — can slowly direct the recollection of the higher category.

Previous studies have considered the gradual activation of attractors by sequences of stimuli [14, 2]. However, in these cases, the order of stimuli is fixed, while we consider the case where it changes for each episode. This problem is closer in concept to using neural networks for encoding and recovering hierarchical information [9, 10]. In these modular networks, the lower level concepts are stored in different modules, and the higher level categories are represented by the activation of all the component modules. Hierarchical networks learn patterns on multiple levels by representing them through specially designed patterns where the common part is retained in the high level concepts [9, 10]. We do not use such patterns.

2 Problem Definition

The system's task is to progressively recognize a set of context inputs presented sequentially in a random order in between irrelevant stimuli.

The network is presented with m different external stimulus sequences in discrete time. Each sequence, S^q, of length n

$$S^q = C_1^q[R^q]C_2^q[R^q]...[R^q]C_p^q[R^q] \quad q = 1, 2, ..., M \qquad (1)$$

starts with a sub-sequence of r patterns – *the context sequence* – comprising p context patterns, C_i^q, interspersed with varying numbers of non-context patterns, indicated by $[R^q]$. The last $[R^q]$ is a sub-sequence of $n - r$ non-context stimuli termed the *regular sequence*. The context patterns, C_i^q are drawn from a set of patterns called the *context set*, $C = \{C_k\}$, and the remaining patterns are drawn from the set, $R = \{R_k\}$. Both C_k, $R_k \in I$, where I is the input space of dimension N_i. For purposes f simulation, patterns in both sets are generated randomly.

A total of n possible contexts are defined, each specified by a unique set of p context patterns drawn randomly without repetition from C. The context sequence for each sequence, S^q, includes context patterns for a unique context, albeit in random order and mixed with non-context patterns. The regular sequence for each S^q is randomly chosen from R.

At the beginning of an episode, as context patterns are presented, each context pattern specifies the correct attractor with increasing certainty until, at the end of the context sequence, the attractor is uniquely identified. THus, the network activity should gradually become confined to the active set of the correct context and to remain confined during the presentation of the regular stimulus.

3 Method

The approach uses a four-layer latent attractor network. The *stimulus layer*, L_S, has N_S neurons that project the input patterns to the *response layer*, L_R. The connections from L_S to L_R are set randomly with probability p_S of connection. Only K_S neurons in the input layer are active at any one time. The *biasing layer* L_B has N_B neurons that receive input from the stimulus layer L_S and project to the response layer. The input connections from L_S to L_B layer are chosen randomly, with a high probability of connection p_{B1}. The output connections from L_B to L_R layer are also chosen randomly with a high probability of connection (p_{B2}).

The response layer, with N_R neurons, also receives a disynaptic recurrent connection through the *intermediary*

layer, L_H with N_H neurons. The latent attractors are stored in the recurrent connections between the L_R and L_H layers. There are M attractors, each comprising two binary patterns, one in layer L_R and the other in layer L_H. The patterns are sparse, with G_R and G_H active neurons, respectively in L_R and L_H. The sets of G_R and G_H neurons that would be active if the attractor were fully manifested are termed the *active sets* of the attractor. The connections between L_R and L_H layers are chosen randomly with probability of connections p_R (L_H to L_R) and p_H (L_R to L_H). The attractors are embedded in the connections by setting the weights according to a clipped binary Hebbian rule first proposed by Willshaw: The connections between neurons active in the two patterns of any attractor are set to high values, while the rest are set to low values [20]. In this way, the M pairs of patterns are set as attractors or fixed points in the in the 2-layer network. The attractors are called latent because they are not allowed to become fully active at any time. The activity in the network is determined as follows. The excitation to a layer L_R neuron, i, at time t is given by:

$$y_i(t) = \sum_{j \in L_S} w_{ij} x_j(t) + g_i(t) \sum_{j \in L_H} w_{ij} z_j(t-1) + g_{bias} \sum_{j \in L_B} w_{ij} u_j(t-1) \qquad (2)$$

where w_{ij} denote connection weights, $x_j(t)$ is the jth bit of the external stimulus patterns at time t, $z_j(t)$ is the output of neuron $j \in L_H$, $u_j(t)$ is the output of neuron $j \in L_B$, $g_i(t)$ is the (modiafible) recurrent gain of neuron i, and g_{bias} is the gain from the biasing layer L_B to the response layer L_R.

The excitation to a layer L_H neuron, i, is given by:

$$y_i(t) = \sum_{j \in L_R} w_{ij} z_j(t) \qquad (3)$$

where $z_j(t)$ is the output of $j \in L_R$.

The input onto a layer L_B neuron i is:

$$y_i(t) = \sum_{j \in L_S} w_{ij} x_j(t) \qquad (4)$$

where $x_j(t)$ is the output of neuron $j \in L_S$.

Firing in both L_R and L_H is competitive: The output of the K_R (K_H) most excited neurons in L_R (L_H) at time t is set to 1, while the rest of the neurons output 0. This corresponds to a K-winner take all competitive firing rule. The values of K_R and K_H respectively are much smaller than G_R and G_H, the respective sizes of the attractors active sets in each layer.

Latent attractors are associated with stimulus sequences as follows: The connections between L_S and L_R layers are modified such that context patterns in each sequence (C^q) stimulate mostly neurons in the active set of the corresponding attractor in the L_R layer. The role of the biasing layer L_B is to sustain the level of activity in candidate latent attractors during the presentation of irrelevant patterns in the context sequence. Each neuron in the L_B layer is associated with one context pattern. When a context pattern is presented at the stimulus layer L_S, the associated neuron in the biasing layer becomes active. The active biasing neuron, in turn, increases the excitation of the L_R neurons in the latent attractors that are associated with the corresponding context pattern. Thus, in between context patterns, the activity in the latent attractors tends to be preserved until a new context pattern is presented at the input. The activity of the biasing neurons is reset after a latent attractor has been fully activated (i.e. at the end of a context sequence). The biasing layer acts essentially as a working memory by sustaining the effect of context patterns until the context sequence is complete.

An important parameter in this network is the recurrent gain, g_i, since it controls the stability of attractors by determining the strength of the recurrent projection to L_R relative to the external stimulus. For a latent attractor to be persistent, neurons in its active set must have a minimum value of g_i [6].

The primary task of the network is to create consistent context-dependent representations for stimuli while maintaining any similarity information between them [3]. Thus, two similar stimuli should produce similar responses if presented in the same context, but different ones if presented in distinct contexts. Retaining similarity information between stimuli is important because it often provides important cues (e.g., indicating spatial proximity for sensory stimuli). We have shown previously that achieving the twin (and apparently incompatible) objectives of preserving stimulus information and sustaining latent attractors requires careful (but not critical) management of the recurrent gain, g_i [5], and also that disynaptic recurrent networks are better able to support this than monosynaptic ones [6].

During the presentation of the context sequence, the network's state should not settle until it has received enough context patterns to uniquely identify the sequence, and should not be disrupted by the intervening irrelevant stimuli. This is achieved in our system through a process we term *incremental competitive positive feedback* [8]. The stability of any attractor in the network is controlled by the recurrent gains, g_i. When g_i are small (relative to the strength of the external stimulus), attractor dynamics is dominated by the impact of the feed-forward stimulus. If the stimulus is selectively associated with particular sets of neurons, these are likelier to win the competition for firing among L_R neurons. If the g_i are large, the recurrent path dominates and the network is forced to choose between

attractors due to competitive firing in L_R and L_H.

In our system, all g_i are set to a small value at the beginning of an episode, so that attractors that are associated with the early context patterns are likely to be activated a bit more than others due to feed-forward association. As the presentation of context patterns proceeds, g_i for neurons that belong to the active sets of attractors with more current activity is increased gradually, priming these attractors for possible persistence if reinforced by subsequent context stimuli. Thus, at each stage, activity is distributed among those attractors that are consistent with the context stimuli received thus far. As each new context stimulus is presented, some of these candidate attractors are reinforced further at the expense of others until, finally, only one is left. When the stimulus is not a context pattern, it causes no significant change in the bias for any attractor, and the biasing neurons keep activity distributed among the attractors as at the previous step.

The equation governing the modulation of recurrent gain is:

$$\hat{g}_i(t) = g_{min} + \frac{g_{max} - g_{min}}{(1 + e^{-\alpha(a_l(t)-\beta)})}$$

$$d_i(t) = \hat{g}_i(t) - g_i(t-1))$$

$$g_i(t) = \begin{cases} \hat{g}_i(t) & if\ |d_i(t)| < \Delta g_{max} \\ g_i(t-1) + \Delta sgn(d_i(t)) g_{max} & else \end{cases} \quad (5)$$

where α is a rate of change parameter, β is an offset parameter, l is the index of the attractor for which i is in the active set, $a_l(t)$ is the total number of active neurons in the L_R active set of attractor l at time t, g_{min} and g_{max} are the minimum and maximum possible values of the recurrent gain. Thus, the gain is $\hat{g}_i(t)$, but the absolute change in gain is bounded by Δg_{max} ($\Delta > 0$).

The modulation of recurrent gain on individual neurons is motivated by several biological considerations:

1. Projections to neurons in most cortical regions are segregated on the dendritic tree, making the selective modulation of gain on input from individual sources quite feasible [16].

2. It is well known that, in the hippocampal region, which is the basis for our model, animals are especially attentive at the beginning of an episode, as indicated by the change in the EEG theta rhythm. This leads to, for example, greater spike synchronization, lower firing latency, and other phenomena [19].

3. In the granule cells of the dentate gyrus, which, we hypothesize, corresponds roughly to our layer L_R, there is both anatomical and physiological evidence [15, 17] of an intricate and highly specific system of excitability modulation based on motivation and attention [?, 19].

Our model represents an attempt to understand how these mechanisms may help support gradual context selection in the hippocampus and artificial neural networks motivated by the hippocampus.

4 Simulation Results

Simulation were done using a two layer latent attractor network with the following parameters: $N_S = 400$, $K_S = 40$, $p_S = 0.4$, $N_B = 20$, $p_{B1} = 0.9$, $p_{B2} = 0.9$, $N_R = 2000$, $G_R = 200$, $K_R = 40$, $p_R = 0.4$, $N_H = 500$, $G_H = 50$, $K_H = 45$, $p_H = 0.8$. There are $M = 10$ attractors embedded in the connections between L_R and L_H layers. The modulation rate for recurrent gain g_i is $\alpha = 0.5$ and $\beta = 33$. The gain of the L_B to L_R projection is $g_{bias} = 6$.

The context set C has 20 distinct patterns, from which 5 context sets are selected. Each C^q consists of $p = 5$ distinct patterns picked randomly without repetition from C. The context patterns in different C^q are not mutually exclusive. Each C^q set of context patterns is associated with a randomly chosen attractor: The connections from L_S to L_R layers are potentiated such that patterns in C^q are associated with the neurons in the active set of the appropriate attractor. Also, each individual context pattern, C_k is associated with a neuron, k, in the biasing layer through Hebbian potentiation of the connections from the stimulus layer L_S and the L_B layer. In turn, each biasing neuron, k, provides excitation to neurons in the active sets of those latent attractors whose context sequences include C_k.

At the beginning of each sequence, S^q, the recurrent gain for all L_R neurons is set to a low value. Depending on how many context groups are simultaneously stimulated by the incoming context patterns from C^q, the activity in L_R and L_H is distributed among the excited attractors. The recurrent gain of neurons in these attractors goes up, while that of other neurons decreases. At the end of a context sequence, only one attractor is consistent with the whole set of context stimuli in C^q, and almost all activity should be concentrated in its active set.

In the first set of simulations, the context patterns are presented in a random order at the network input layer interleaved with irrelevant patterns. The total length of the context sequence is fixed to $r = 20$, and the number of context patterns is $p = 5$, but the position of the remaining irrelevant patterns is not fixed. Figure 1 shows the result of a single network simulation, when the context sequences are presented at the input. Each graph represents the normalized activity within the active set of an attractor pattern in the L_R layer. It can be seen that, for each context sequence, the activity in only one of the attractors goes up steadily. In all other attractors the activity might

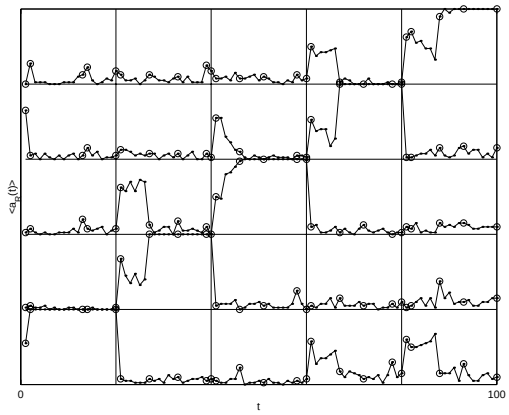

Figure 1: The activity level in the selected attractors in the L_R layer with respect to time. Every $r = 20$ time steps a different context sequence starts. The activity is normalized with to K_R. The time steps when context patterns are presented is denoted with circles.

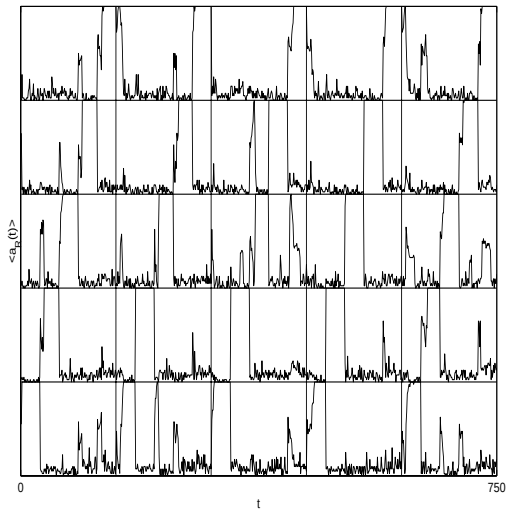

Figure 2: Several (five) repeats of the run in Figure 1, with different context pattern order in each set, each time.Each context sequence is followed by a regular sequence of 10 patterns.

increase for a few time steps, but it finally shuts down. In between consecutive context patterns, the activity is spread approximately equally between the candidate attractors. Figure 2 shows the results when the simulation is repeated with the same 5 contexts sequences but with the context patterns presented in different order each time. Each context sequence is followed by 10 regular patterns. It is clear that the activity remains confined within the chosen attractor even though the regular patterns have no association with any attractor. It can be seen that sometimes a wrong attractor almost wins the competition (the spikes) in the middle of a context sequence, but it is finally shut down.

In the second set of simulation results, the overlap between two context sequences is varied, keeping the order of the context patterns the same (i.e. the context sequences start with the coinciding context patterns) and the interval between consecutive context patterns is the same. Figure 3 shows the mean activity in two attractors as compared to the mean activity in the non-context attractors. For an overlap of two patterns, (Figure 3 (a)), as well for an overlap of three patterns (Figure 3 (b)), the mean activity in the context attractors goes up slowly, and in between context patterns it remains almost at the same level. Eventually, the correct attractor is selected. Figure 4 shows a sample simulation result obtained with one network. The attractors which show activity share have the first two context patterns. Again, in between the first two context patterns the activity level tends to remain constant.

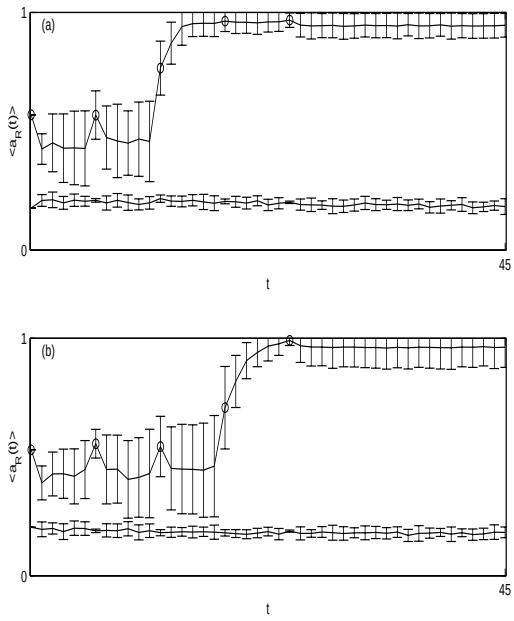

Figure 3: Both gaphs show the mean activity level in two context attractors (upper curve) and the mean activity level in non-context attractors (lower curve)during the presentation of a stimulus seuqence. The mean is taken over five different networks and five different presentations of the same context patterns. Figure (a) corresponds to an overlap of two context patterns between the context sets of the two selected attractors, while in Figure (b) there is an overlap of three context patterns. Each context sequence ($r = 25, p = 5$) has a constant inter context pattern interval of five regular patterns, and is followed by 20 regular patterns.

5 Conclusions

We have proposed a mechanism by which an attractor in a latent attractor network can be activated progressively by a set of inputs presented sequentially in a random order, rather than by a single cue. In the interval between consecutive relevant inputs, the network sees a variable number of irrelevant patterns. The system is able to select and ac-

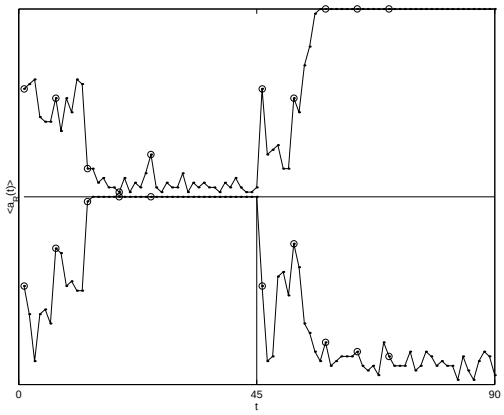

Figure 4: The two graphs represent the total activity in two attractors, with the first two context patterns identical. The moment when context patterns are presented to the input is signaled by circles. There are five regular patterns in between the context ones, and each context sequence is followed by 20 regular patterns.

tivate the right attractor progressively, even though, each individual input pattern can be associated with more than one attractor. The system can overcome the disrupting effect of the irrelevant patterns by trying to maintain the same state of attractor participation until a new relevant pattern is encoutered. Importantly, the response is invariant to the order in which the context patterns within a set are presented.

Acknowledgement: This research was supported by grant no. IBN 980664 from NSF and by a grant from the University Research Council, University of Cincinnati.

References

[1] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. on Neural Networks*, vol. Vol. 5, No. 2, pp. 157–166, 1994.

[2] G. Bradski, G.A. Carpenter and S. Grossberg. STORE working memory networks for storage ad recall of arbitrary temporal sequences. *Biological Cybernetics* 71:469–480, 1994.

[3] S. Doboli, A.A. Minai and P.J. Best. Generating smooth context-dependent representations. *Proc. of IJCNN'1999*

[4] S. Doboli, A.A. Minai, and P.J. Best. A latent attractors model of context-selection in the dentate gyrus-hilus system. *Neurocomputing* 26-27:671–676, 1999.

[5] S. Doboli, A.A. Minai and P.J. Best. Latent attractors: a model for context-dependence place representations in the hippocampus. *Neural Computation* 12:1009–1043, 2000.

[6] S. Doboli and A.A. Minai. Network capacity for network attractor computation. *Proc. IJCNN'2000* 222-228, 2000.

[7] S. Doboli, A.A. Minai, and P.J. Best, A comparison of context-dependent hippocampal place codes in 1-layer and 2-layer recurrent networks, *Neurocomputing*, 3-33:353–358, 2000.

[8] S. Doboli, A.A. Minai, Progressive attractor selection in latent attractor networks, *Proc. IJCNN'2001*, 2001.

[9] D.R.C. Dominguez. Information capacity of a hierarchical neural network. *Phys. Rev. E* 58:4811–4815, 1998.

[10] V.S. Dotsenko. Hierarchical model of memory. *Physica A*, 410–415, 1986.

[11] J.L. Elman, "Finding structure in time," *Cognitive Science.*, vol. 14, pp. 179–211, 1990.

[12] P. Frasconi and M. Gori, "Computational capabilities of local-feedback recurrent networks acting as finite-state machines," *IEEE Trans. on Neural Networks*, vol. Vol. 7, No. 6, pp. 1521–1525, 1996.

[13] C.L. Giles, C.B. Miller, D. Chen, H.H. Chen, G.Z. Sun, and Y.C. Lee, "Learning and extracting finite state automata with second-order recurrent neural networks," *Neural Computation*, vol. 4, pp. 393–405, 1992.

[14] S. Grossberg and C. W. Myers. The resonant dynamics of speech perception: Interword integration and duration-dependent backward effects. *Psychological Review*, in press.

[15] Z.-S. Han, E.H. Buhl, Z. Lörinczi, and P. Somogyi, A high degree of spatial selectivity in the axonal and dendritic domains of physiologically identified local-circuit neurons in the dentate gyrus of the rat hippocampus. *Eur. J. Neurosci.* 5, 395–410, 1993.

[16] M.E. Hasselmo, E. Schnell, and E. Barkai. Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in hippocampal region CA3. *J. Neurosci.*, 15:5249–5262, 1995.

[17] M.B. Jackson and H.E. Scharfman, Positive feedback from hilar mossy cells to granule cells in the dentate gyrus revealed by voltage-sensitive dye and microelectrode recording. *J. Neurophysiol.* 76, 601–616, 1996.

[18] A.A. Minai and P.J. Best. Encoding spatial context: A hypothesis on the function of the dentate gyrus-hilus system. *Proc. of IJCNN'1998* 587-598, 1998.

[19] E.I. Moser. Altered inhibition of granule cells during spatial learning in an exploration task. *J. Neurosci.* 16:1247-1259.

[20] D. Willshaw, O.P. Buneman and H.C. Longuet-Higgins. Non-holographic associative memory. *Nature* 222:960-962.