Memory Dynamics in Attractor Networks with Saliency Weights

Huajin Tang
htang@i2r.a-star.edu.sg
Institute for Infocomm Research, Agency for Science Technology and Research, Singapore 138632

Haizhou Li
hli@i2r.a-star.edu.sg
Institute for Infocomm Research, Agency for Science Technology and Research, Singapore 138632, and Department of Computer Science and Statistics, University of Eastern Finland, 80101 Joensuu, Finland

Rui Yan
ryan@i2r.a-star.edu.sg
Institute for Infocomm Research, Agency for Science Technology and Research, Singapore 138632

Memory is a fundamental part of computational systems like the human brain. Theoretical models identify memories as attractors of neural network activity patterns based on the theory that attractor (recurrent) neural networks are able to capture some crucial characteristics of memory, such as encoding, storage, retrieval, and long-term and working memory. In such networks, long-term storage of the memory patterns is enabled by synaptic strengths that are adjusted according to some activity-dependent plasticity mechanisms (of which the most widely recognized is the Hebbian rule) such that the attractors of the network dynamics represent the stored memories. Most of previous studies on associative memory are focused on Hopfield-like binary networks, and the learned patterns are often assumed to be uncorrelated in a way that minimal interactions between memories are facilitated. In this letter, we restrict our attention to a more biological plausible attractor network model and study the neuronal representations of correlated patterns. We have examined the role of saliency weights in memory dynamics. Our results demonstrate that the retrieval process of the memorized patterns is characterized by the saliency distribution, which affects the landscape of the attractors. We have established the conditions that the network state converges to unique memory and multiple memories. The analytical result also holds for other cases for variable coding levels and nonbinary levels, indicating a general property emerging from correlated memories. Our results

confirmed the advantage of computing with graded-response neurons over binary neurons (i.e., reducing of spurious states). It was also found that the nonuniform saliency distribution can contribute to disappearance of spurious states when they exit.

1 Introduction

The brain is able to encode and store representations of the external world through memory. Electrophysiological recordings in various brain regions have suggested that human memory involving the neocortex for short-term memory and the hippocampus for long-term memory (Jensen & Lisman, 2004; Bird & Burgess, 2008) has similar structure to computer memories. However, memory mechanisms are fundamentally different: while classic computer memories rely on static approach, biological memories are based on firing activities of neurons driven by internal dynamics such that memories are represented as stable network activity states called attractors (Hopfield, 1982; Amit, Gutfreund, & Sompolinsky, 1985a).

Attractor networks have been central to neuronal models of memory for more than three decades for both experimentalists and theorists (Amit, Gutfreund, & Sompolinsky, 1985b; Dayan & Willshaw, 1991; Chechik, Meilijson, & Ruppin, 2001; Pantic, Torres, Kappen, & Gielen, 2002; Amit & Mongillo, 2003; Poucet & Save, 2005; Tsodyks, 2005), as they possess such an appealing feature that the underlying attractor dynamics theory can provide a unified description of several aspects of memory, including encoding, storage, retrieval, and long-term and short-term memory. The hypothesis of attractor dynamics is supported by the persistent activity observed in the neocortex and hippocampus during memory experiments (Miyashita, 1988; Ericson & Desimone, 1999; Bakker, Kirwan, Miller, & Stark, 1999). It was also shown that hippocampal region CA3, characterized by heavy recurrent connections and modifiability, could be an anatomical substrate where the attractor networks reside (Treves & Rolls, 1994; Lisman, 1999; Moser, Kropff, & Moser, 2008).

Attractor networks perform a sparse encoding of memories by mapping a continuous input space to a sparse output space, which is composed of a discrete set of attractors. When a stimulus pattern close to a stored pattern is presented to the system, the network states are drawn by the intrinsic dynamics toward the attractor that corresponds to the memorized pattern. One important feature imprinted with the attractor networks is that memory retrieval depends on the attractor landscape, which reflects the capability of pattern completion (recall the original input from a degraded version) and pattern separation (without mixing up the other stored memories) of high-level functions. The formation of the attractor landscape is achieved by Hebbian-type synaptic modifications (Hebb, 1949) in which each synapse is involved in the storage of multiple items. Obviously this common synaptic representation implies interactions between memories stored in the same network. The associative networks studied in previous work are successful in capturing some fundamental characteristics of associative memory, such as memory capacity and associative capability (Amit et al., 1985a; Mueller & Herz, 1999; Chechik et al., 2001).

Most previous studies on associative neural network models suffer from two limitations. First, the models are restricted to networks with binary (0/1 or $-1/1$) neurons or graded response (e.g., sigmoid functional) neurons, which cannot capture the features that biological neurons are seldom firing with saturations. The steady states of such networks can also arise from saturation, thus leading to spurious attractor states. Second, in these models, the patterns are often assumed to be uncorrelated so that the interactions in the resulting memories are minimal. Indeed, uncorrelated patterns are not unique to the memory. Some direct evidence has emerged recently to show that attractor dynamics was used to encode and retrieve memories of correlated shapes about their environments (Wills, Lever, Cacucci, Burgess, & O'Keefe, 2005; Leutgeb et al., 2005). Wills et al. (2005) recorded the firing activity of hippocampal place cells in the brains of freely moving rats exposed to a square or circular environment. When the rats explored the circle and the square environment initially, the responses of the ensemble activity of hippocampal cells were different so as to differentiate between the two environments. To see how the neuronal representations changed when the animals were in the intermediate environments, the rats were then tested in a set of environments of intermediate shapes between the circle and the square. Surprisingly, most cells fired in a pattern that was either circle-like or square-like, and no gradual change in the representations was observed; the neurons abruptly and simultaneously switched from one activity pattern to the other at the same midpoint between the circle and the square. Wills et al. explained their results as a strong support of the presence of two distinct attractors in the hippocampus network corresponding to the circular and square environments. Many of the hippocampus neurons changed their representations sharply and coherently at a certain position along the morph sequence, where basins of attraction of the two attractors meet. Similar experiments were performed in Leutgeb et al. (2005), but with different results. Instead of observing two distinct activity patterns, Leutgeb et al. reported that the neuronal representations of the subsequent environments in the morph sequence gradually change. The results indicate the flexibility of memory representation in attractor networks and put forth new challenges to provide a unified description using attractor networks' modality.

The theoretical work of Blumenfeld, Preminger, Sagi, and Tsodyks (2006) was the first attempt to tackle such challenges; they showed that attractor network models allow memory representations of correlated stimuli that depend on learning order. In their study, the Hopfield network (Hopfield, 1982) was used to study the attractor dynamics of long-term memory, where the states of neurons have binary values ($+1$ for active neuron and -1

for inactive neuron). A novel form of facilitated learning was proposed to explain that different learning protocols could be the reason for the incongruous experimental observations.

In this study, we aim to extend the work on associative memory by using more biologically plausible network models and exploring the role of saliency weights on the memory dynamics for the attractor networks. Our analytical result is consistent with the previous study that when network inputs are correlated, this mechanism results in overassociations or splitting of attractors. The model predicts that memory representations should be sensitive to the saliency weights formed by the learning mechanism, compatible with recent electrophysiological experiments on hippocampal place cells. Section 2 introduces the associative memory model. Section 3 presents the stability conditions of the attractor networks. Section 4 presents the memory dynamics analysis and the analytical solution for the correlated binary patterns and gives numerical examples. More general cases for variable coding level and nonbinary patterns and disappearance of spurious states are studied in sections 5 and 6. Finally, these results are discussed in section 7.

2 The Model

Hebbian synaptic plasticity has been the major paradigm for studying memory and self-organization in computational neuroscience. Within the associative memory framework, many Hebbian learning rules have been suggested in the neural network literature. However, the relationship between the input patterns and the landscape of attractors by using the Hebbian regime has not been established, and no procedure can robustly translate an arbitrary specification of an attractor landscape into a set of weights. One of the challenges is that knowledge in the network is distributed over connections; each connection participates in specifying multiple attractors. Therefore, developing neural networks as a practical memory device is still an intriguing problem that continues to attract interest (Zemel & Mozer, 2001; Siegelmann, 2008).

In the following, we describe an analog associative memory model, namely, the linear-threshold associative (LTA) model, which consists of N neurons with the linear-threshold activation function $\sigma(x) = \max(0, x)$. The ith neuron updates its firing state x_i at time t according to the total synaptic input it receives from the network via synaptic connection strengths w_{ij},

$$\dot{x}_i(t) = -x_i(t) + \sigma\left(\sum_{j=1}^{N} w_{ij} x_j(t) + h_i\right), \tag{2.1}$$

where h_i is its external input.

Through the Hebbian synaptic modification mechanism, correlations within patterns to be memorized are encoded in the synaptic weights.

By this procedure, multiple patterns can be implemented as fixed-point attractors of the network dynamics. Starting from an initial state close to one of the stored attractors, the system dynamics relaxes to this attractor and thus retrieves the stored pattern. Associative memory storage in a dynamic system requires the existence of multiple attractors. The attractors are imprinted by the Hebbian rule.

The synaptic strength w_{ij} from neuron jth (presynaptic) to neuron ith (postsynaptic) is determined by a general additive synaptic learning rule that depends on the P stored memory patterns ξ_μ,

$$w_{ij} = \sum_{\mu=1}^{P} \left(\xi_i^\mu - c\right)\left(\xi_j^\mu - c\right), \tag{2.2}$$

which represents the Hebbian learning rule for associative memory networks, where c is the coding level.[1] The $\xi^\mu (\mu = 1, \ldots, P)$ denote P memory-generating patterns to be stored and later to be retrieved by the network. To enable different patterns with different saliencies in forming the synaptic weights, we introduce an extended model where patterns are stored in the network with a variable saliency factor:

$$w_{ij} = \sum_{\mu} s(\mu)\left(\xi_i^\mu - c\right)\left(\xi_j^\mu - c\right). \tag{2.3}$$

The stimuli ξ^μ (μ being the index of the corresponding pattern) to be stored as memory are encoded as network activity patterns. Then the memories can be retrieved by providing an initializing input and allowing the network to evolve autonomously by updating neurons' activity states according to equation 2.1.

In the original Hopfield model, the connections acquire the following values after all of the patterns are learned:

$$w_{ij} = \frac{1}{N} \sum_{\mu} \xi_i^\mu \xi_j^\mu .$$

We will show that this original formulation does not allow representations of multiple memories for the correlated patterns.

In equation 2.3, the Hebb-like terms $(\xi_i^\mu - c)(\xi_j^\mu - c)$ are modulated by a corresponding saliency factor $s(\mu)$. This approach is inspired by the fact

[1] Throughout this letter, the coding level is defined as the mean level of activity of the network (Treves, 1990a; Pantic et al., 2002). In some previous work it is also defined as a fraction of firing neurons (e.g., Chechik et al., 2001). For binary 0/1 neurons, they are equivalent; otherwise they are not.

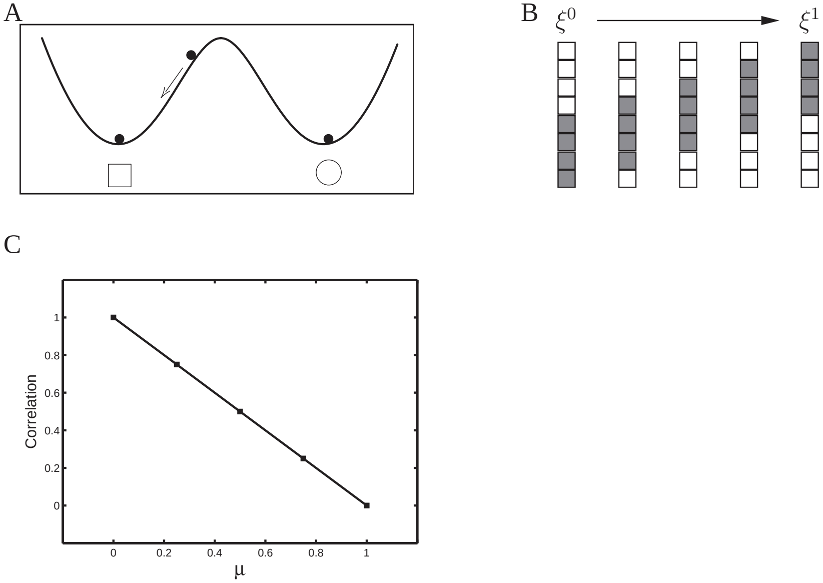

Figure 1: Example of morphed environments in experiments and morphed patterns in our model. (A) Animals explored a set of morphed environments between square and circular shapes, and the neuronal representations of the hippocampal cells switched to either of the attractor states corresponding to the square or the circle (adapted from Figure C of Poucet and Save, 2005). (B) A sequence of patterns was used to model the morphed shapes of environments. ξ^μ (μ from 0 to 1) indicating the correlation between a pair of patterns (ξ^μ and ξ^v) decreased gradually from 1 to 0 as the distance between them increases. In each pattern, half of the neurons are $+1$ (shaded boxes) and half are 0 (empty boxes). In this figure, eight neurons with five patterns are shown. (C) The correlation is decreased gradually from 1 to 0 as the distance between patterns increased.

that the patterns presented for learning may acquire different saliency captured by factor $s(\mu)$, and a large value implies a large saliency on the weights.

We will use the model to store a sequence of gradually changing patterns that mimics the set of morph stimuli used in the experiments described above. We will consider a set of binary patterns, ξ^μ, with the index μ now representing the position of the pattern in the morph sequence ($0 \leq \mu \leq 1$; see Figure 1 for an example). The sequence is constructed by first choosing the source pattern (ξ^0) and the target pattern ξ^1, which are uncorrelated ($\xi^{0^\top} \xi^1 = 0$); in every step of morphing, we gradually change the state of a fixed number of neurons such that the target pattern is obtained at the end. In the standard analysis of attractor dynamics, identifying the entire set of stored patterns crucially depends on the assumption that the patterns are uncorrelated:

$$\frac{1}{N}\sum_i \xi_i^{\mu}\xi_i^{v} = 0$$

for every $\mu \neq v$, such that the interference between different patterns during memory retrieval is minimal. However, biological memories are not restricted to such assumptions as they also memorize correlated patterns as evidenced by experiments (Wills et al., 2005; Leutgeb et al., 2005). Thus, it is natural to consider neuronal representations with correlated ones.

For the morphed patterns constructed by the above procedure, the correlation between a pair of patterns (ξ^{μ} and ξ^{v}) decreases gradually from 1 to 0 as the distance between them increases:

$$\frac{1}{N}\sum_i \xi_i^{\mu}\xi_i^{v} = 1 - |\mu - v|.$$

This results in a much stronger interference between the patterns, and it could lead to undesired network behaviors when stored memories are being retrieved.

3 Stability Conditions

The model consists of unsaturating responses of neurons; thus, its stability has to be maintained by suitable interactions between the ensemble of neurons. The network dynamics has been analyzed in a series of theoretical studies (Hahnloser, 1998; Wersing, Beyn, & Ritter, 2001; Yi, Tan, & Lee, 2003; Tang, Tan, & Teoh, 2006; Tang, Tan, & Zhang, 2005), where conditions for fixed-point attractors and limit cycles were examined. It is revealed that the stabilities (monostability and multistability) are only dependent on the synaptic strengths.

Define a new connection matrix W^+ with entries

$$w_{ij}^{+} = \delta_{ij} w_{ij} + (1 - \delta_{ij}) \max(0, w_{ij}), \tag{3.1}$$

where δ_{ij} is defined by $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. The connections W^+ can be considered the result of performing the linear thresholding on the off-diagonal elements of W.

If the connection weights satisfy

$$w_{ii} + \sum_{j \neq i} w_{ij}^{+} < 1 \tag{3.2}$$

for all i, then the dynamics of the model is bounded (Wersing et al., 2001).

For any i, let the set P^- denote the set of the off-diagonal elements that are nonpositive, that is, $P^- = \{j \mid w_{ij} \leq 0\}$. We can get the following criterion: if

$$\begin{cases} r = 1 - w_{ii} - \sum_{j \neq i} w_{ij}^{+} > 0 \\ \sum_{j \in P^-} w_{ij} < -r \end{cases} \tag{3.3}$$

for all i, then the network is multistable, that is, multiple attractor states can coexist.

Equation 3.3 suggests that $\lambda_{\max}\{W^+\} < 1 < \lambda_{\max}\{W\}$. It is noted that the synapse $w_{ij} = \sum_\mu s(\mu)\xi_i^\mu \xi_j^\mu$ always lets $\lambda_{\max}\{W^+\} \geq \lambda_{\max}\{W\}$; therefore, the synapse takes the following form,

$$w_{ij} = \frac{1}{N} \sum_{\mu} s(\mu)\big(\xi_i^\mu - c\big)\big(\xi_j^\mu - c\big), \tag{3.4}$$

for the N-units network.

4 Memory Dynamics with Saliency Weights

In the previous section, the stability conditions do not make any predictions on the memory retrieval process. Now we will show the memory dynamics and examine the role of saliency weights on the emergence of attractor states. We follow a procedure similar to that of Blumenfeld et al. (2006), based on the analysis of the overlap (a variable measuring the similarity between the network activity states and the stored patterns).

4.1 Analysis. The overlap between the instantaneous network state x and the memorized pattern ξ^μ is now defined in terms of the coding level:

$$m^\mu = \frac{1}{N} \sum_{j=1}^{N} \big(\xi_j^\mu - c\big) x_j. \tag{4.1}$$

The overlap provides insight into the network dynamics and measures the similarity of the actual state and the stored patterns.

The dynamics of the states of the network are

$$\begin{aligned} \dot{x}_i(t) &= -x_i(t) + \sigma\left(\sum_j w_{ij} x_j + h_i\right) \\ &= -x_i(t) + \sigma\left(\sum_j \sum_v \frac{1}{N} s(v)\big(\xi_i^v - c\big)\big(\xi_j^v - c\big) x_j + h_i\right). \end{aligned} \tag{4.2}$$

The network states can be described through the overlap:

$$\begin{aligned}\dot{m}^{\mu}(t) &= \frac{1}{N}\sum_{i}\left(\xi_i^{\mu}-c\right)\left(-x_i+\sigma\left(\sum_{j}\sum_{\nu}\frac{1}{N}s(\nu)\left(\xi_i^{\nu}-c\right)\left(\xi_j^{\nu}-c\right)x_j+h_i\right)\right)\\ &= -m^{\mu}(t)+\frac{1}{N}\sum_{i}\left(\xi_i^{\mu}-c\right)\sigma\left(\sum_{\nu}\left(\xi_i^{\nu}-c\right)s(\nu)m^{\nu}+h_i\right).\end{aligned}$$

In the above formulation, since h_i does not affect the stability, we ignore this term to get

$$\dot{m}^{\mu}(t) = -m^{\mu}(t)+\frac{1}{N}\sum_{i}\left(\xi_i^{\mu}-c\right)\sigma\left(\sum_{\nu}\left(\xi_i^{\nu}-c\right)s(\nu)m^{\nu}\right). \tag{4.3}$$

The network dynamics converges to a stable steady state, ξ^{μ^*}, which is an attractor of the network. In contrast to the Hopfield network of binary neurons, where the retrieved memory is identified by the overlapping value $m^{\mu^*}=1$ (Blumenfeld et al., 2006), the retrieved memory of our model is identified by the maximal overlapping value over all the patterns, that is, $\mu^* = \arg\max_{\mu} m^{\mu}$, since the neuronal responses are unsaturated to binary states.

Next, we define

$$I_i^{\mu} = \left(\xi_i^{\mu}-c\right)\sigma\left(\sum_{\nu}\left(\xi_i^{\nu}-c\right)s(\nu)m^{\nu}\right) \tag{4.4}$$

and divide the summation over the neuron index i into four parts,

$$\dot{m}^{\mu}(t) = -m^{\mu}(t)+\frac{1}{N}\sum_{i\in C_0^{\mu+}} I_i^{\mu}+\frac{1}{N}\sum_{i\in C_0^{\mu-}} I_i^{\mu}+\frac{1}{N}\sum_{i\in C_1^{\mu+}} I_i^{\mu}+\frac{1}{N}\sum_{i\in C_1^{\mu-}} I_i^{\mu}, \tag{4.5}$$

with the sets of indexes $C_0^{\mu+}$, $C_0^{\mu-}$, $C_1^{\mu+}$, $C_1^{\mu-}$ defined below. For each neuron, we assign an index, $0 < \rho_i \leq 1$, which denotes the location in the morph sequence where the component ξ_i begins to change its value (from 0 to 1 or vice versa). First, we consider that $\xi_i^{\nu}=0$ for $\nu < \rho_i$ and $\xi_i^{\nu}=1$ for $\nu \geq \rho_i$. The set $C_0^{\mu+}$ is defined to be the set of neuron indexes i satisfying $\rho_i > \mu$ and $C_0^{\mu-}$ to be the set of neuron indexes i satisfying $\rho_i \leq \mu$. Then we consider that $\xi_i^{\nu}=1$ for $\nu < \rho_i$ and $\xi_i^{\nu}=0$ for $\nu \geq \rho_i$. We similarly define the set $C_1^{\mu+}$ for neurons satisfying $\rho_i > \mu$ and $C_1^{\mu-}$ for neurons satisfying

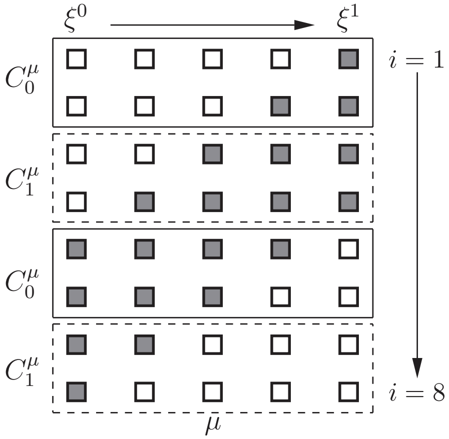

Figure 2: Definitions of the neuron index sets $C_0^{\mu+}, C_0^{\mu-}, C_1^{\mu+}, C_1^{\mu-}$ for $\mu = 0.5$.

$\rho_i \le \mu$. Mathematically, it is written as

$$C_0^{\mu+} = \left\{ i \mid \xi_i^v = 0 \text{ if } v < \rho_i \text{ and } 1 \text{ if } v \ge \rho_i;\ \rho_i > \mu \right\}$$

$$C_0^{\mu-} = \left\{ i \mid \xi_i^v = 0 \text{ if } v < \rho_i \text{ and } 1 \text{ if } v \ge \rho_i;\ \rho_i \le \mu \right\}$$

$$C_1^{\mu+} = \left\{ i \mid \xi_i^v = 1 \text{ if } v < \rho_i \text{ and } 0 \text{ if } v \ge \rho_i;\ \rho_i > \mu \right\}$$

$$C_1^{\mu-} = \left\{ i \mid \xi_i^v = 1 \text{ if } v < \rho_i \text{ and } 0 \text{ if } v \ge \rho_i;\ \rho_i \le \mu \right\}.$$

Figure 2 illustrates the different sets of neuron indices for overlapping the pattern $\mu = 0.5$.

In the limit of large number of patterns, the retrieved memory index μ^* can be found by solving the following equation (the derivation procedure is given in the appendix):

$$\int_0^{\mu^*} s(\mu) \left(\left(\mu^* - \frac{1}{2} \right)^2 - (\mu - \mu^*)^2 + \frac{1}{4} \right) d\mu$$
$$= \int_{\mu^*}^1 s(\mu) \left(\left(\mu^* - \frac{1}{2} \right)^2 - (\mu - \mu^*)^2 + \frac{1}{4} \right) d\mu. \tag{4.6}$$

For some special choice of the saliency distribution, equation 4.6 can be solved explicitly. For the uniform saliency distribution, $s(\mu) \equiv s$ for all μ, the above formulation is reduced to

$$\int_0^{\mu^*} \left(\left(\mu^* - \frac{1}{2} \right)^2 - (\mu - \mu^*)^2 + \frac{1}{4} \right) d\mu$$
$$= \int_{\mu^*}^1 \left(\left(\mu^* - \frac{1}{2} \right)^2 - (\mu - \mu^*)^2 + \frac{1}{4} \right) d\mu. \tag{4.7}$$

Solving the integrals, we get the single solution $\mu^* = 0.5$. If, however, the saliences are biased toward the source and the target of the morph sequence ($\mu = 0$ and $\mu = 1$, respectively), the network can acquire two attractors corresponding to the first and the last patterns, respectively. For example, if $s(\mu) = 6(\mu - 0.5)^2$, then there are three solutions to equation 4.6: two solutions, $\mu^* = 0.5 + \frac{\sqrt{4\sqrt{10}-5}}{6}(\approx 0.039)$ and $\mu^* = 0.5 - \frac{\sqrt{4\sqrt{10}-5}}{6}(\approx 0.961)$, corresponding to two stable attractors, and one solution, $\mu^* = 0.5$, corresponding to an unstable attractor. These results demonstrate that different distributions of saliency induce different attractor landscapes.

4.2 Uniform Saliency Yields a Single Attractor. First, we show how the uniform saliency weights affect the attractor states. A number of 17 morphed patterns formed as in Figure 1 are encoded into the synaptic weights by using equation 2.3, and the network consists of 32 neurons. In the simulation, we set $s(\mu) = 0.6$ and external inputs $b = \frac{1}{P}\sum_{\mu}\xi^{\mu}$. The network evolves from random initial configurations according to a parallel updating dynamics. It is shown that a single attractor is reached (see Figure 3A), which corresponds to the midpoint in the morph sequence $\mu = 0.5$. In accordance with the theoretical prediction made in the previous section, the pattern $\xi^{0.5}$ takes the maximal overlapping value (Figures 3B and 3C) among all the morph patterns when the network stabilizes.

4.3 Nonuniform Saliency Allows Multiple Attractors. Next, we show the effects of nonuniform scaling of Hebbian terms on the attractor landscape. In the simulation, we choose a saliency weights distribution $s(\mu) = 6(\mu - 0.5))^2$. The network starts from random initial states. When stabilized, the activities of the network converge to two distinct attractor states (see Figure 4A). In the two attractor states, the pattern $\mu = 0$ (denoted by the square in the solid line) and pattern $\mu = 1$ (denoted by the circle in the dashed line) takes the maximal overlapping values, respectively (see Figure 4B), indicating that the retrieved memories bias toward one of the edge patterns depending on which is favored by the initial state (see Figures 4C and 4D). This result also verifies the prediction of the developed theory.

The above results are compatible with that of the theoretical work (Blumenfeld et al., 2006) and also consistent with the experiments of Wills et al. (2005) who interpreted their results as evidence for the presence of two distinct attractors corresponding to the circular and square environments. Instead of observing two distinct place-dependent activity patterns, Leutgeb et al. (2005) reported that the similarity between representations of the subsequent environments in the sequence gradually decreased. In their theoretical contributions, Blumenfeld et al. (2006) explained that different phenomena could result from different learning protocols.

5 Different Coding Schemes

Our analysis in the section 4 leads to a solution similar to that of Blumenfeld et al. (2006) based on the assumption that the stored memory patterns are

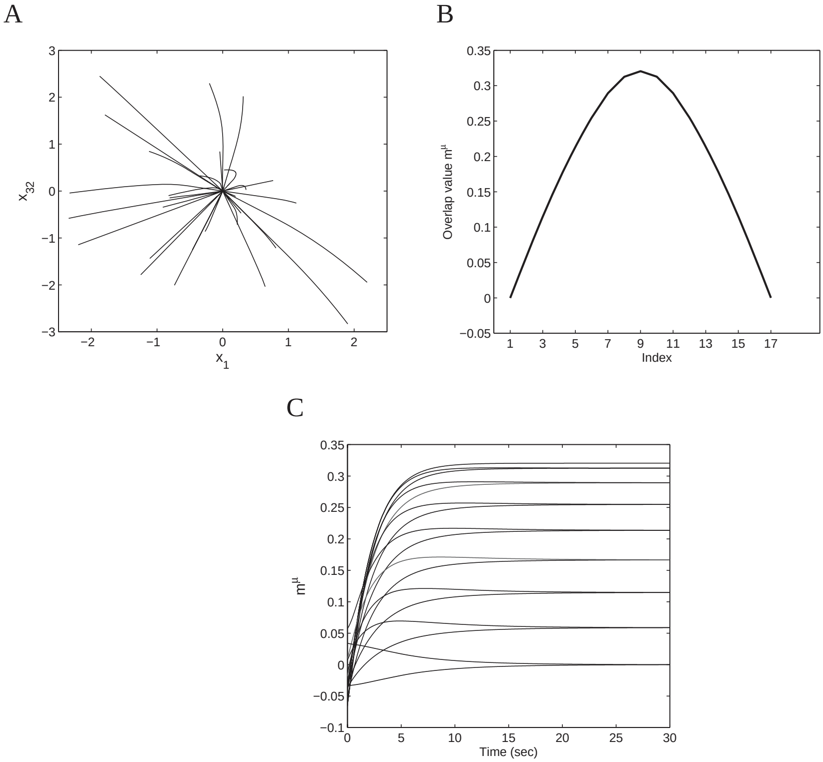

Figure 3: A single stable attractor state stemming from the uniform distribution of saliency weights $s(\mu) = 0.6$. (A) The phase portrait of the single attractor in the $x_1 - x_{32}$ plane. (B) The overlap values of the attractor show the retrieved pattern at the midpoint of the sequence (the ninth index, $\mu^* = 0.5$). (C) Time evolution of the overlapping $m^\mu(t)$ in B.

binary vectors with $c = \frac{1}{2}$, which confirmed that the saliency weight (scaling of Hebbian terms) has the same influence on the formation of the attractor landscape in binary neural networks and linear threshold neural networks. Now an interesting question arises: Does the solution also hold for other cases where $c \neq \frac{1}{2}$?

Since real neurons behave very differently from binary units, their activity can be approximated better by an analog than a binary variable. The biological plausibility and computational advantages of the graded-response networks with threshold linear units have been elucidated in previous analytical studies (Treves, 1990a, 1990b; Treves & Rolls, 1991; Roudi & Treves, 2003). Though with linear threshold neurons, the memory patterns encoded in synaptic weights can be taken to be binary vectors, as we did in the above

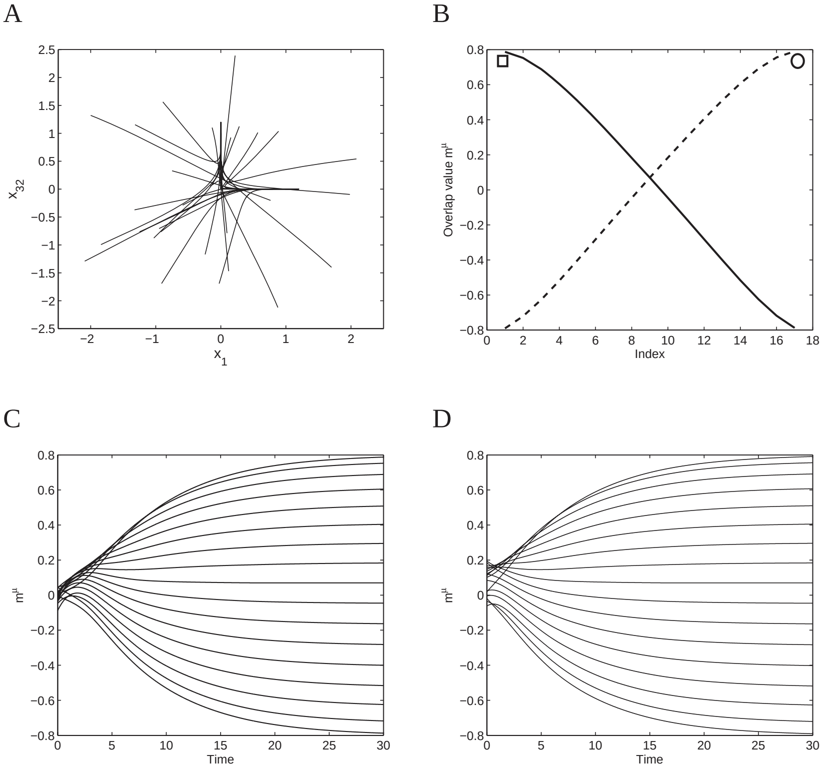

Figure 4: Two attractor states arising from the saliency weights $s(\mu) = 6(\mu - 0.5)^2$. (A) The phase portrait of the two attractors in $x_1 - x_{32}$ plane. (B) The normalized overlapping values of the corresponding two attractors (in the solid line, $\mu^* = 0$, representing the "square" environment, in the dashed line: $\mu^* = 1$ representing the "circular" environment of Figure 1A). (C, D) Time evolution of the overlapping $m^\mu(t)$ of the "square" and "circular" attractor states in B, respectively.

studies, they can also be taken to be drawn from a distribution with several discrete activity values or from a continuous distribution. Hence, the second question arises: Does the conclusion also hold for graded-response networks storing nonbinary memory patterns?

In this section, we attempt to answer the two questions through numerical studies. To quantify the effects of different coding schemes on the associative performance of attractor networks, a number of different coding schemes in Treves and Rolls (1991) are studied. We adopt these coding schemes; that is, the activity η of each neuron follows the probability

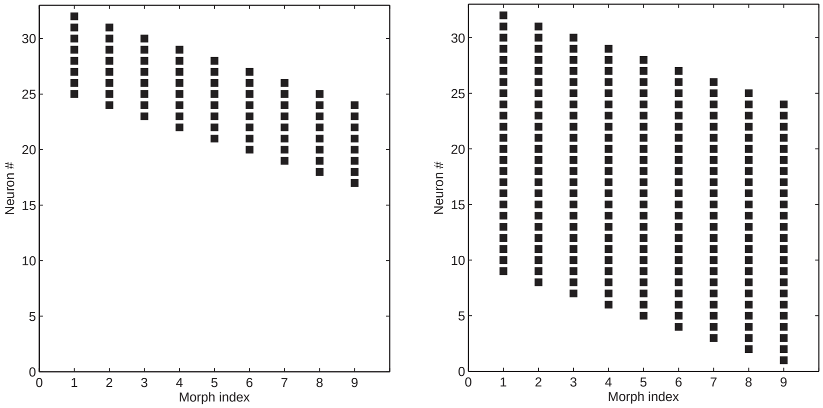

Figure 5: Raster plots of the morphed binary patterns ($N = 32$, $P = 9$) with variable coding levels: $c = \frac{1}{4}$ (left) and $c = \frac{3}{4}$ (right). The neural activities are shown in gray scale (the maximal activity value 1 is shown in black, and the minimal activity 0 is in white).

distribution function $p(\eta)$:

$$\begin{aligned} p(\eta) &= (1-c)\delta(\eta) + c\delta(\eta - 1), \qquad \text{(binary)} \\ p(\eta) &= \left(1 - \frac{4c}{3}\right)\delta(\eta) + c\delta\left(\eta - \frac{1}{2}\right) + \frac{c}{3}\delta\left(\eta - \frac{3}{2}\right), \quad \text{(ternary)} \qquad (5.1) \\ p(\eta) &= (1-2c)\delta(\eta) + 4ce^{-2\eta}, \qquad \text{(exponential)}, \end{aligned}$$

where $\delta(x)$ is the Dirac's function.

As Treves and Rolls (1991) noted, the ternary distribution offers a good prototypical example to probe into the features emerging with nonbinary structures, and the exponential distribution is biologically meaningful because it is consistent with the continuous distribution demonstrated by experimental data.

5.1 Variable Coding Levels.

As in the extensions of Hopfield model, we thus allow for the parameter c to differ from the value $c = 0.5$. Figure 5 demonstrates the morphed binary memory patterns for different coding levels ($c = 0.25$ and 0.75), where nine patterns are encoded into a network with 32 neurons. In simulations, the network receives the same uniform background input b.

We vary different values of c to examine retrieval performance and find that the saliency weights have a consistent effect on the attractor landscapes. The simulation results are given in Figure 6 for uniform saliency

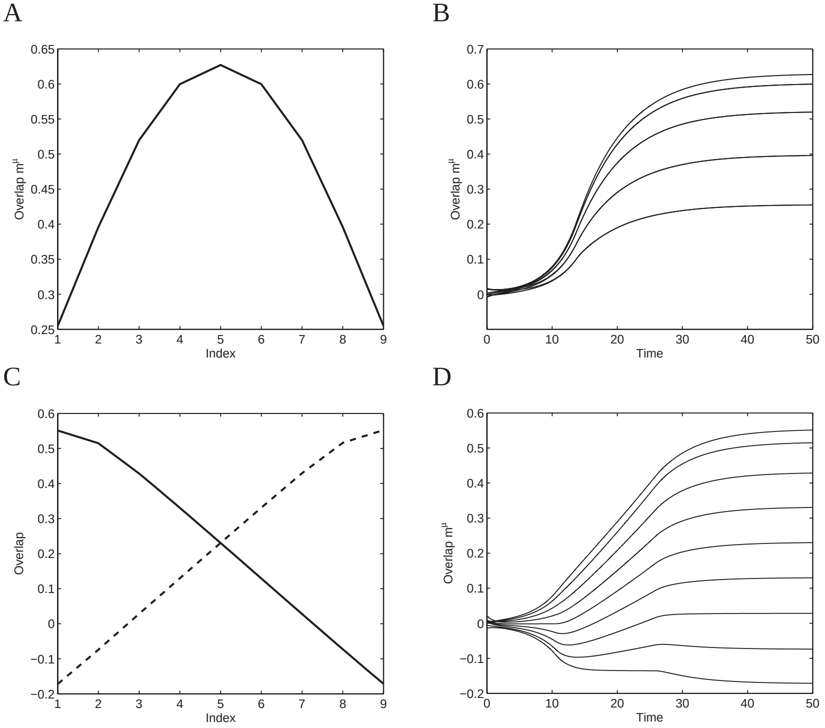

Figure 6: Overlap values for $c = 0.25$. (A, B) A uniform saliency distribution, uniform saliency $s(\mu) = 1.4$, is applied, and the retrieved memory is the middle of the sequence. (C, D) A nonuniform saliency distribution $s(\mu) = 14(\mu - 0.5)^2$ is applied, and the retrieval tends to be either of the edge patterns of the stored sequence.

and nonuniform saliency distribution. For $c = 0.75$, the results are given in Figure 7 for both cases.

5.2 Nonbinary Memory Patterns. To understand the effect of increasingly structured memories, it is necessary to consider nonbinary structured memories. The ternary patterns, representing the simplest nonbinary structure, become a natural choice for studying the features that emerged with nonbinary patterns (Treves, 1990a). Hence, we are prompted to use the ternary patterns in our numerical studies.

A sequence of correlated patterns is generated according to the ternary distribution in equation 5.1. Figure 8 shows the raster plot of nine correlated ternary patterns with the coding level $c = 0.25$, where the neural activities are indicated by gray scale.

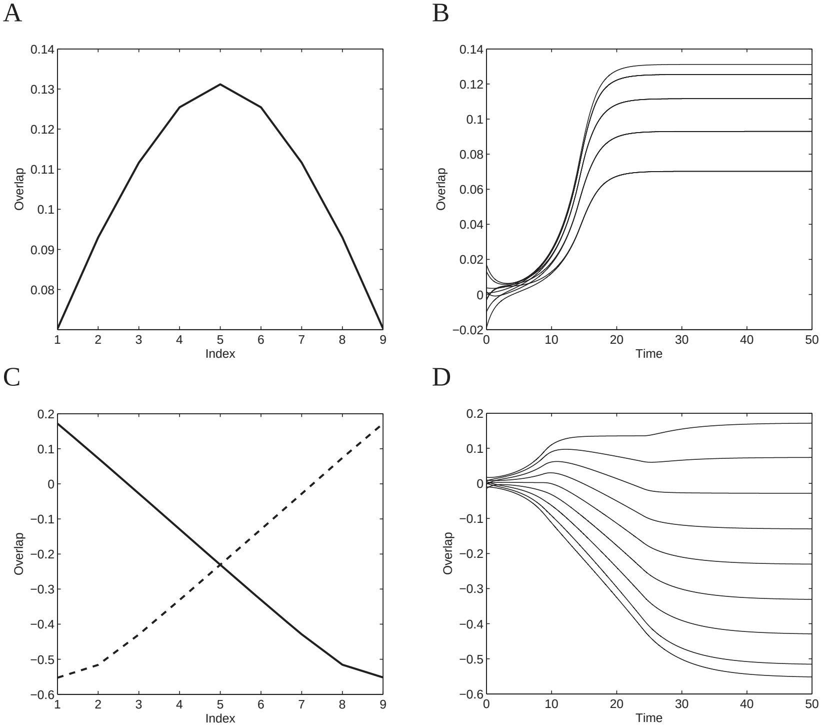

Figure 7: Overlap values for $c = 0.75$. (A, B) A uniform saliency distribution, uniform saliency $s(\mu) = 1.4$, is applied, and the retrieved memory is the middle of the sequence. (C, D) A nonuniform saliency distribution $s(\mu) = 14(\mu - 0.5)^2$ is applied, and the retrieval tends to be either of the edge patterns of the stored sequence.

In the numerical simulations, the network evolves from random initial configurations that are positively correlated with the stored patterns. We also vary the coding levels. The effects of uniform saliency and nonuniform saliency weights on the attractor landscape are investigated, and the results are in accordance with that of the binary cases. It is demonstrated that the uniform saliency weight induces a single attractor in the middle of the morphed sequence, while the nonuniform saliency drives the network to retrieve one of the two extreme patterns. Figure 9 shows an example of such results for $c = 0.25$.

6 Disappearance of Spurious States

In attractor network models, individual memory (a pattern of activity distribution) is encoded as an attractor in an energy landscape. Whether the

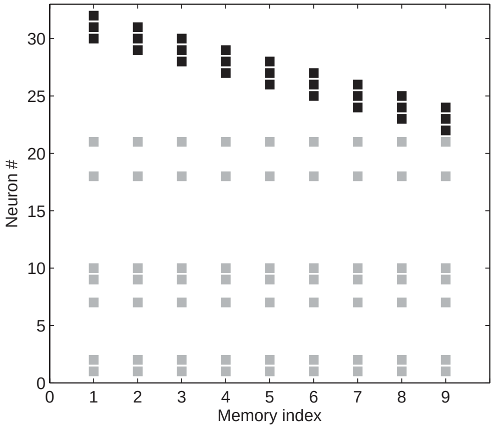

Figure 8: A raster plot of the neural activities of all neurons for the correlated ternary memory patterns ($N = 32$, $P = 9, c = 0.25$). The neural activities with the ternary coding are shown in gray scale according to their strengths.

stored memory can be retrieved safely without being trapped by dynamical hurdles depends critically on two aspects: the smoothness of the energy landscape and the existence of other attractor states (namely, spurious states). Waugh, Marcus, and Westervelt (1990) showed analytically that neural networks with analog neurons have a computational advantage over their binary-neuron counterparts, as analog-valued neurons contribute to smoothing the energy landscape and thus greatly reduce the number of spurious states. The question of mixture states, related to the other aspects of the dynamical hurdles, was studied analytically in Roudi and Treves (2003), where the mean field solutions and their stability conditions were developed by considering an associative network with threshold-linear units. It was revealed that symmetric n-mixture states, $n = 2, 3, \dots,$ are almost never stable, and only with a binary coding scheme can a limited region of the parameter space be found in which either 2- or 3-mixtures are stable. This property demonstrates further the advantage of computation with graded-response neurons, as the stability region of the spurious states is eliminated by nonbinary coding schemes that are naturally endowed with the network of graded-response units.

It remains interesting to understand when the same binary coding scheme is used whether the stability of mixture states in a graded-response network is more restricted than that in binary valued network, and whether the saliency weights can contribute to smooth the attractor landscape (or reduce the mixture states).

6.1 Networks of Binary Neurons and Graded-Response Neurons. First, we consider 2-mixture states in a network of binary neurons and a

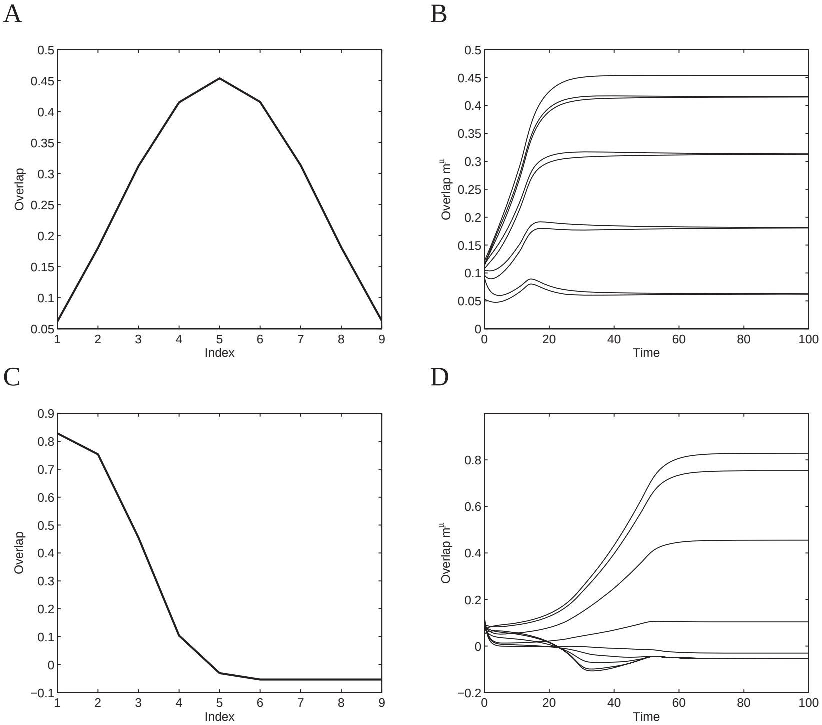

Figure 9: Overlapping values for retrieval of ternary patterns with $c = 0.25$. (A, B) With a saliency distribution $s(\mu) = 2$, the overlap of the middle pattern dominates the others. (C, D) With $s(\mu) = 15(\mu - 0.5)^2$, the overlap of the first pattern dominates the others.

network of graded-response (linear-threshold) neurons, respectively. The encoded memory patterns are generated using the same binary coding scheme described in equation 5.1: $N = 1000$, $P = 3$. For large N, the patterns are uncorrelated as $\frac{1}{N}\xi^{\mu\top}\xi^{v} = 0$, $\mu \neq v$.

The binary network, a modification from that of Blumenfeld et al. (2006), enabling 0/1 coding, is described by

$$x_i(t+1) = \text{sign}\left(\sum_{j=1}^{N} w_{ij}x_j(t)\right), \tag{6.1}$$

where sign(·) is defined as $\text{sign}(x) = 1$ if $x > 0$ and $\text{sign}(x) = 0$ if $x \leq 0$. The memories are encoded by the synaptic mechanism, equation 3.4, which takes into account the coding level, unlike what was used in Blumenfeld et al. (2006).

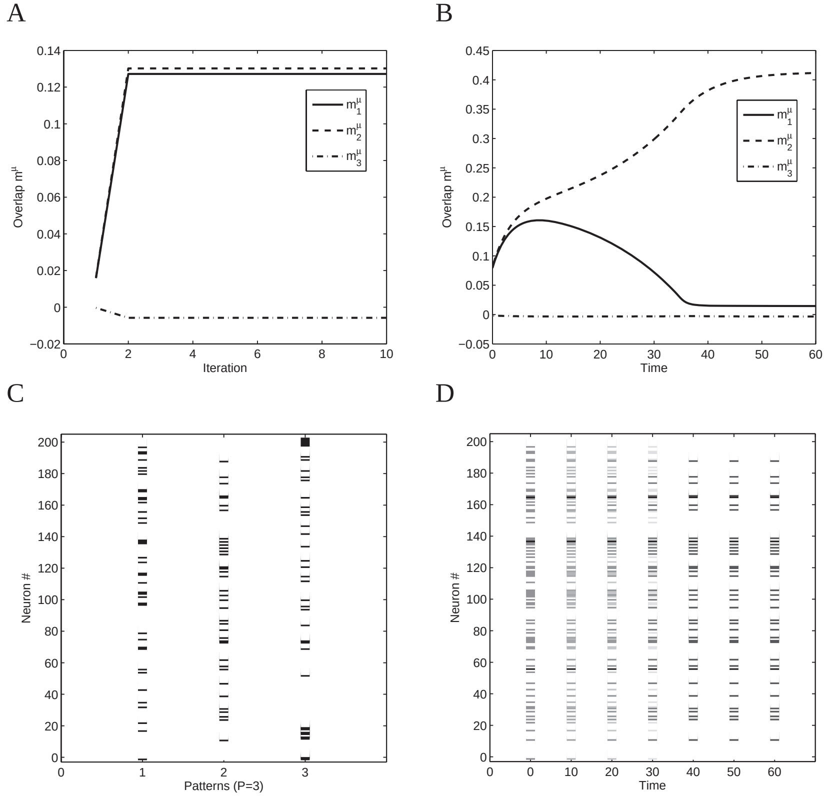

Figure 10: Computer simulation of 2-mixture states for the network of linear neurons $N = 1000$, $P = 3$, $s(\mu) = 6.5$. The initial state was correlated equally with two stored patterns with $c = 0.2$. The network of binary neurons evolves toward a spurious state, whereas the network of linear-threshold neurons is reaching the corresponding attractor (B). This result demonstrates that the same saliency results in different attractor landscape in the two networks. (C) The raster plot of the activities of the first 200 neurons for $p = 1, 2, 3$. (D) The raster plot of activities of the first 200 neurons during retrieval, at every 100 steps (finally the second pattern is retrieved).

We run the networks from the same initial configuration, which is correlated equally with two patterns. The simulation results are shown in Figure 10. For the same saliency distribution, the binary network typically evolves toward one of the spurious states (see Figure 10, left). In contrast, the spurious state disappears in the graded-response network, as shown in Figure 10 (right), in which one of the overlaps tends to dominate, reaching the corresponding attractor, whereas the other one tends to zero.

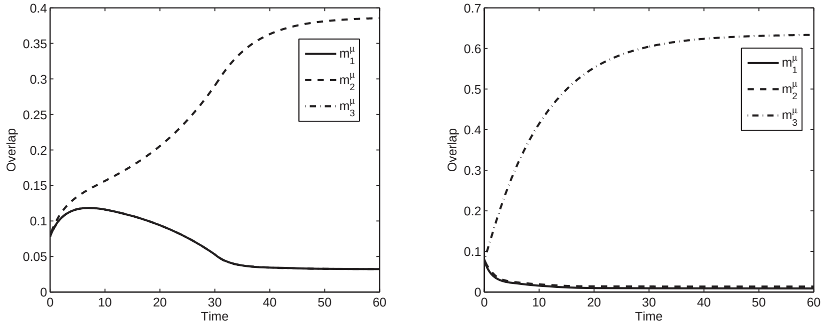

Figure 11: Computer simulation of 3-mixture states for the network of threshold-linear neurons, $N = 1000$, $P = 3$, $c = 0.2$. (Left) With $s(\mu) = 6.5$, two overlaps grow together. (Right) With $s(\mu) = 11(\mu - 0.5)^2$, one overlap dominates the others, indicating the corresponding attractor is reached. The initial state is correlated equally with three stored patterns.

This result illustrates that given the same saliency distribution, different attractor landscapes can be expected in networks with binary neurons and linear-threshold neurons. It is also in agreement with previous findings that the free-energy landscape is smoothed by using analog neurons (Waugh et al., 1990; Roudi & Treves, 2003).

Next, we investigated the effect of different saliency distribution on mixture states in the linear-threshold network. Our computer simulations find that when the mixture states exist, a nonuniform saliency distribution contributes to eliminating them, as shown in Figure 11. In the simulations, the 3-mixture states are studied, and the initial states are correlated equally with three stored patterns. With a uniform distribution, a mixture state is reached (two overlaps tend to grow; see Figure 11, left), while with a nonuniform distribution, a single overlap dominates the others (see Figure 11, right) implying that the corresponding memory pattern is retrieved.

6.2 Related Work. An associative model with linear-threshold (or threshold-linear) neurons was proposed first in Treves (1990a). In the model, the activity of each unit is determined by its input through the linear-threshold function

$$x_i = g\,\sigma(h_i - T_{thr}), \tag{6.2}$$

where g is a gain parameter and T_{thr} is a threshold, where the input to each unit i takes the form

$$h_i = \sum_{j \neq i} J_{ij} x_i + \sum_{v} s^v \frac{(\xi_i^v - c)}{c} + b\left(\frac{1}{N}\sum_{j} x_j\right), \tag{6.3}$$

where the first two terms are related to the memory encoding and partial cue retrieval, respectively.

Unlike model 2.1, in Treves's model, excitatory and inhibitory neurons are represented separately with different functionality. The memory patterns are superimposed onto the excitatory-excitatory connections through the Hebb rule:[2]

$$J_{ij} = \frac{1}{Nc^2} \sum_{\mu=1}^{P} \left(\xi_i^\mu - c\right)\left(\xi_j^\mu - c\right). \tag{6.4}$$

A b term is added to regulate the activity of the network to a desired level ($\frac{1}{N}\sum_i x_i = \frac{1}{N}\sum_i x_i^2 = c$) at any moment. The b terms plays a role of uniform inhibition, depending on the average network activity. Biological plausibility and computational advantages have been discussed extensively in Treves (1990a, 1990b) and Treves and Rolls (1991) and recently in Roudi and Treves (2003), where a theoretical framework for analyzing the storage capacity and associative performance has been established by studying the mean field solutions of the attractor states.

Through extensive numerical studies on more general cases, including variable coding levels and nonbinary patterns, it is shown that our solution is not restricted to the particular case for gradually changing binary memories. In fact, reshaping of the attractor landscape through a different saliency distribution is a robust property emerging from correlated memories stored in the graded-response neural networks. Our results also confirmed the computational advantages of graded-response neurons (i.e., eliminating mixture states). Interestingly, a nonuniform saliency can also contribute to the disappearance of spurious states when they exist.

From the modeling aspect, Treves's model, equation 6.3, presents a more attractive feature in encoding the memories of pattern activities, being more flexible than the LTA model, which was also studied in an application to analog associative memories (Tang et al., 2006). As in the former, the neural activities are modulated separately by the b term, an approximation of the lumped inhibition effects, and the synaptic weights work as information encoding only. In contrast, the LTA model mixed the roles of memory encoding and activity (stability) control into a single term of synaptic weights, without distinguishing excitatory and inhibitory contributions. Though given the difference in modeling local inputs, similar memory retrieval properties have been observed for the two models (see Figures 10 and 1 in Roudi & Treves, 2003) on uncorrelated memories. Hence, our results can be considered complementary to such studies on associative memories on the networks with graded-response neurons.

[2] A different overlap formulation was used there: $x^\mu = \frac{1}{Nc}\sum_{i=1}^{N} \eta_i^\mu x - \bar{x}$, $\bar{x}$ is the mean of the network states. It is noted that $m^\mu = \frac{1}{N}\sum_{j=1}^{N}(\xi_j^\mu - c)x_j = c(\frac{\xi^\top x}{Nc} - \bar{x}) = c \cdot x^\mu$.

7 Discussion

In this contribution we studied an attractor network with a biologically realistic model for the associative memory of correlated patterns. Through theoretical analysis, we examined the mechanisms for encoding and retrieval memories and the dynamics of memory representations. It was shown that saliency weights can dramatically affect the resulting memory representations. The model identified memories as attractor states of network with weights adhering to a Hebb-like process of long-term synaptic modifications and retrieved via internal dynamics of patterned network activity. The results are in line with the recent experimental observations by Wills et al. (2005) and extend the previous theoretical contributions by Blumenfeld et al. (2006). This model allows the merging and splitting of attractors, which leads to overassocications (several memory merging into one) or complete associations (retrieving all memory items), and supports the flexibility of the attractor representations through novelty facilitated learning mechanisms proposed by Blumenfeld et al. (2006). Based on their novelty-facilitated learning scheme, the incongruous findings in Wills et al. (2005) and Leutgeb et al. (2005) could be formulated as the results of different learning order.

Our model, representing a large class of recurrent network models, will provide a convenient framework for studying long-term memory and its relation to pattern separation and pattern completion. The model is also useful for investigating the design of the weights for different input patterns and the effects of learning protocol on memory representations. For the system in the regime of activity-dependent plasticity, if given enough exposure time and a stable sensory environment, it would create a stable attractor state for each item with an equal attractive basin. The novelty-facilitated or history-dependent learning strategy allows reshaping the profile of saliency weights, and thus strengthens or weakens the impact of individual memorized items. It would be expected that the attractive basins are enlarged or shrink, resulting in the merging or splitting of memories.

Appendix: Analysis of Attractor States

The following analysis gives the procedure for finding attractor states through the overlapping dynamics $m^{\mu}(t)$.

For $i \in C_0^{\mu+}$, it holds that

$$\begin{aligned} I_i^{\mu} &= -\frac{1}{2}\sigma \left(\sum_{\nu<\rho_i} \left(\xi_i^{\nu} - c\right) s(\nu) m^{\nu} + \sum_{\nu\geq\rho_i} \left(\xi_i^{\nu} - c\right) s(\nu) m^{\nu} \right) \\ &= -\frac{1}{2}\sigma \left(-\frac{1}{2} \sum_{\nu<\rho_i} s(\nu) m^{\nu} + \frac{1}{2} \sum_{\nu\geq\rho_i} s(\nu) m^{\nu} \right), \end{aligned} \tag{A.1}$$

which used the fact that $\xi_i^\mu - c = -\frac{1}{2}$. Then

$$\frac{1}{N}\sum_{i\in C_0^{\mu+}} I_i = \frac{1}{N}\sum_{i\in C_0^{\mu+}} -\frac{1}{2}\sigma\left(-\frac{1}{2}\sum_{\nu<\rho_i} s(\nu)m^\nu + \frac{1}{2}\sum_{\nu\geq\rho_i} s(\nu)m^\nu\right). \tag{A.2}$$

At every morph step, we updated $\frac{N}{P}$ neurons; then the summation takes the form

$$\frac{1}{N}\sum_{i\in C_0^{\mu+}} I_i = \frac{1}{P}\sum_{\rho_i>\mu} -\frac{1}{2}\sigma\left(-\frac{1}{2}\sum_{\nu<\rho_i} s(\nu)m^\nu + \frac{1}{2}\sum_{\nu\geq\rho_i} s(\nu)m^\nu\right). \tag{A.3}$$

Using the assumption of large P and ρ_i is a uniformly distributed variable, we replace the sum over the set $C_0^{\mu+}$ with an integral to get

$$\frac{1}{N}\sum_{i\in C_0^{\mu+}} I_i = -\frac{1}{4}\int_\mu^1 d\rho\ \sigma\left(-\int_0^\rho s(\nu)m^\nu d\nu + \int_\rho^1 s(\nu)m^\nu d\nu\right). \tag{A.4}$$

For $i\in C_0^{\mu-}$, it holds that

$$\begin{aligned} I_i^\mu &= \frac{1}{2}\sigma\left(\sum_{\nu<\rho_i}\left(\xi_i^\nu - c\right)s(\nu)m^\nu + \sum_{\nu\geq\rho_i}\left(\xi_i^\nu - c\right)s(\nu)m^\nu\right) \\ &= \frac{1}{2}\sigma\left(-\frac{1}{2}\sum_{\nu<\rho_i} s(\nu)m^\nu + \frac{1}{2}\sum_{\nu\geq\rho_i} s(\nu)m^\nu\right), \end{aligned} \tag{A.5}$$

which used the fact that $\xi_i^\mu - c = \frac{1}{2}$. We replace the sum over $C_0^{\mu-}$ with an integral to get

$$\begin{aligned} \frac{1}{N}\sum_{i\in C_0^{\mu-}} I_i &= \frac{1}{P}\sum_{\rho_i\leq\mu}\frac{1}{2}\sigma\left(-\frac{1}{2}\sum_{\nu<\rho_i} s(\nu)m^\nu + \frac{1}{2}\sum_{\nu\geq\rho_i} s(\nu)m^\nu\right) \\ &= \frac{1}{4}\int_0^\mu d\rho\ \sigma\left(-\int_0^\rho s(\nu)m^\nu d\nu + \int_\rho^1 s(\nu)m^\nu d\nu\right). \end{aligned} \tag{A.6}$$

Analogously, we calculate the sums for the other two sets $C_1^{\mu+}$ and $C_1^{\mu-}$, respectively:

$$\begin{aligned} \frac{1}{N}\sum_{i\in C_1^{\mu+}} I_i^\mu &= \frac{1}{P}\sum_{\rho_i>\mu}\frac{1}{2}\sigma\left(\frac{1}{2}\sum_{\nu<\rho_i} s(\nu)m^\nu - \frac{1}{2}\sum_{\nu\geq\rho_i} s(\nu)m^\nu\right) \\ &= \frac{1}{4}\int_\mu^1 d\rho\ \sigma\left(\int_0^\rho s(\nu)m^\nu d\nu - \int_\rho^1 s(\nu)m^\nu d\nu\right). \end{aligned} \tag{A.7}$$

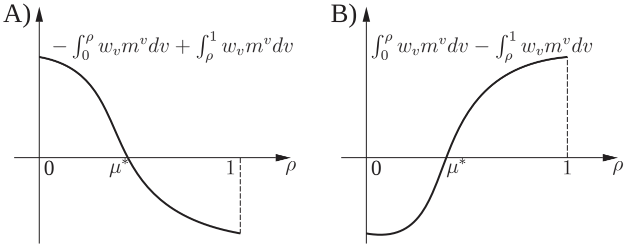

Figure 12: The expression $-\int_0^\rho w_v m^v(t)dv + \int_\rho^1 w_v m^v(t)dv$ and its counterpart.

For $i \in C_1^{\mu-}$,

$$\frac{1}{N}\sum_{i \in C_1^{\mu-}} I_i^\mu = \frac{1}{P}\sum_{\rho_i \le \mu} -\frac{1}{2}\sigma\left(\frac{1}{2}\sum_{v<\rho_i} s(v)m^v - \frac{1}{2}\sum_{v \ge \rho_i} s(v)m^v\right)$$
$$= -\frac{1}{4}\int_0^\mu d\rho\, \sigma\left(\int_0^\rho s(v)m^v dv - \int_\rho^1 s(v)m^v dv\right). \tag{A.8}$$

Substituting the above four parts into equation 4.5, we can get

$$\begin{aligned}\dot{m}^\mu(t) = -m^\mu(t) &- \frac{1}{4}\int_\mu^1 d\rho\, \sigma\left(-\int_0^\rho s(v)m^v dv + \int_\rho^1 s(v)m^v dv\right)\\ &+\frac{1}{4}\int_0^\mu d\rho\, \sigma\left(-\int_0^\rho s(v)m^v dv + \int_\rho^1 s(v)m^v dv\right)\\ &+\frac{1}{4}\int_\mu^1 d\rho\, \sigma\left(\int_0^\rho s(v)m^v dv - \int_\rho^1 s(v)m^v dv\right)\\ &-\frac{1}{4}\int_0^\mu d\rho\, \sigma\left(\int_0^\rho s(v)m^v dv - \int_\rho^1 s(v)m^v dv\right).\end{aligned}$$

We assume that the expression

$$-\int_0^\rho w_v m^v(t)dv + \int_\rho^1 w_v m^v(t)dv$$

crosses zero at a unique point $\rho = \mu^*$ such that

$$\int_0^{\mu^*} s(v)m^v(t)dv = \int_{\mu^*}^1 s(v)m^v(t)dv, \tag{A.9}$$

as illustrated in Figure 12.

Then by using the variable μ^*, we can simplify the equation $\dot{m}^{\mu}(t)$ to get

$$\begin{aligned}\dot{m}^{\mu}(t) = & -m^{\mu}(t) - \frac{1}{2}\int_{\mu}^{1} d\rho\, \sigma\left(\int_{\rho}^{\mu^*} s(v)m^{v}dv\right) \\ & +\frac{1}{2}\int_{0}^{\mu} d\rho\, \sigma\left(\int_{\rho}^{\mu^*} s(v)m^{v}dv\right) \\ & +\frac{1}{2}\int_{\mu}^{1} d\rho\, \sigma\left(\int_{\mu^*}^{\rho} s(v)m^{v}dv\right) \\ & -\frac{1}{2}\int_{0}^{\mu} d\rho\, \sigma\left(\int_{\mu^*}^{\rho} s(v)m^{v}dv\right).\end{aligned}$$

The above damped dynamics converges to a steady state:

$$\begin{aligned}m^{\mu} = & -\frac{1}{2}\int_{\mu}^{1} d\rho\, \sigma\left(\int_{\rho}^{\mu^*} s(v)m^{v}dv\right) + \frac{1}{2}\int_{0}^{\mu} d\rho\, \sigma\left(\int_{\rho}^{\mu^*} s(v)m^{v}dv\right) \\ & +\frac{1}{2}\int_{\mu}^{1} d\rho\, \sigma\left(\int_{\mu^*}^{\rho} s(v)m^{v}dv\right) - \frac{1}{2}\int_{0}^{\mu} d\rho\, \sigma\left(\int_{\mu^*}^{\rho} s(v)m^{v}dv\right).\end{aligned} \tag{A.10}$$

If $\mu < \mu^*$, by applying the central limit theory and the property of $\sigma(\cdot)$, equation A.10 is reduced to

$$\begin{aligned}m^{\mu} = & s(\eta)m^{\mu}(\eta)\left(-\frac{1}{2}\int_{\mu}^{\mu^*} d\rho \int_{\rho}^{\mu^*} dv\right. \\ & \left. +\frac{1}{2}\int_{\mu}^{\mu^*} d\rho \int_{\rho}^{\mu^*} dv + \frac{1}{2}\int_{\mu^*}^{1} d\rho \int_{\mu^*}^{\rho} dv\right),\end{aligned}$$

where $0 < \eta < 1$. We calculate the integrals to get

$$m^{\mu} = k\left(-\frac{(\mu^* - \mu)^2}{2} + \frac{\mu^{*2}}{4} + \frac{(1-\mu^*)^2}{4}\right), \tag{A.11}$$

where k is a scaling constant absorbing $s(\eta)m^{\mu}(\eta)$ and μ^* is a solution satisfying equation A.9.

If $\mu \geq \mu^*$, it is derived from (see equation A.10)

$$\begin{aligned}m^{\mu} = & s(\eta)m^{\mu}(\eta)\left(\frac{1}{2}\int_{0}^{\mu^*} d\rho \int_{\rho}^{\mu^*} dv + \frac{1}{2}\int_{\mu}^{1} d\rho \int_{\mu^*}^{\rho} dv\right. \\ & \left. -\frac{1}{2}\int_{\mu^*}^{\mu} d\rho \int_{\mu^*}^{\rho} dv\right),\end{aligned}$$

and calculating the integral yields the same expression as equation A.11.

From formulation A.11, it can be seen that m^{μ} is maximal for $\mu = \mu^*$. Therefore, the attractor states can be described by ξ^{μ^*} such that

$$\int_0^{\mu^*} s(v)\left(\left(\mu^* - \frac{1}{2}\right)^2 - (\mu - \mu^*)^2 + \frac{1}{4}\right) dv$$
$$= \int_{\mu^*}^{1} s(v)\left(\left(\mu^* - \frac{1}{2}\right)^2 - (\mu - \mu^*)^2 + \frac{1}{4}\right) dv, \qquad \text{(A.12)}$$

which follows from equations A.9 and equation A.11. This gives equation 4.6.

For the simple case that $s(\mu)$ is a constant, the solution μ^* can be solved by

$$\int_0^{\mu^*} \left(\left(\mu^* - \frac{1}{2}\right)^2 - (\mu - \mu^*)^2 + \frac{1}{4}\right) dv$$
$$= \int_{\mu^*}^{1} \left(\left(\mu^* - \frac{1}{2}\right)^2 - (\mu - \mu^*)^2 + \frac{1}{4}\right) dv. \qquad \text{(A.13)}$$

We calculate the above integral equation to get

$$\left(\mu^* - \frac{1}{2}\right)^3 = 0,$$

which has a unique solution at $\mu^* = \frac{1}{2}$. Therefore, if $s(v)$ is constant, the state of the network always converges to the memory $\xi^{0.5}$.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful comments and constructive suggestions.

References

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985a). Spin glass models of neural networks. *Phys. Rev. A., 32*(2), 1007–1018.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985b). Storing infinite number of patterns in a spin glass model for neural networks. *Phys. Rev. Lett., 55*(4), 1530–1533.

Amit, D. J., & Mongillo, G. (2003). Spike-driven synaptic dynamics generating working memory states. *Neural Computation, 15*, 565–596.

Bakker, A., Kirwan, C. B., Miller, M., & Stark, C. E. L. (1999). Pattern separation in the human hippocampal CA3 and dentate gyrus. *Science, 319*, 1640–1642.

Bird, C. M., & Burgess, N. (2008). The hippocampus and memory: Insights from spatial processing. *Nature Reviews Neuroscience, 9*(3), 182–194.

Blumenfeld, B., Preminger, S., Sagi, D., & Tsodyks, M. (2006). Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. *Neuron, 52*, 383–394.

Chechik, G., Meilijson, I., & Ruppin, E. (2001). Effective neuronal learning with ineffective Hebbian learning rules. *Neural Computation, 13*, 817–840.

Dayan, P., & Willshaw, D. J. (1991). Optimizing synaptic learning rules in linear associative memories. *Biol. Cyber., 65*(4), 253–265.

Ericson, C. A., & Desimone, R. (1999). Responses of macaque perirhinal neurons during and after visual stimulus association learning. *J. Neurosci., 19*, 10404–10416.

Hahnloser, R. H. R. (1998). On the piecewise analysis of networks of linear threshold neurons. *Neural Networks, 11*, 691–697.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA, 79*, 2554–2558.

Jensen, O., & Lisman, J. E. (2004). Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends in Neuroscience, 28*(2), 67–72.

Leutgeb, J. K., Leutgeb, S., Treves, A., Meyer, R., Barnes, C. A., McNaughton, B. L., et al. (2005). Progressive transformation of hippocampal neuronal representations in morphed environments. *Neuron, 48*, 345–358.

Lisman, J. E. (1999). Relating hippocampal circuitry to function: Recall of memory sequences by reciprocal dentate-CA3 interactions. *Neuron, 22*, 233–242.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature, 335*, 817–820.

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci., 31*, 69–89.

Mueller, R., & Herz, A. V. M. (1999). Content-addressable memory with spiking neurons. *Physical Review E, 59*(3), 3330–3338.

Pantic, L., Torres, J. J., Kappen, H. J., & Gielen, S. C. A. M. (2002). Associative memory with dynamic synapses. *Neural Computation, 14*, 2903–2923.

Poucet, B., & Save, E. (2005). Attractors in memory. *Science, 308*, 799–800.

Roudi, Y., & Treves, A. (2003). Disappearance of spurious states in analog associative memories. *Physical Review E, 67*, 041906.

Siegelmann, H. (2008). Analog-symbolic memory that tracks via reconsolidation. *Physica D, 237*, 1207–1214.

Tang, H., Tan, K. C., & Teoh, E. J. (2006). Dynamics analysis and analog associative memory of networks with LT neurons. *IEEE Trans. on Neural Networks, 17*(2), 409–418.

Tang, H., Tan, K. C., & Zhang, W. (2005). Analysis of cyclic dynamics for networks of linear threshold neurons. *Neural Computation, 17*, 97–114.

Treves, A. (1990a). Graded-response neurons and information encoding in autoassociative memories. *Physical Review A, 42*(4), 2418–2430.

Treves, A. (1990b). Threshold-linear formal neurons in auto-associative nets. *J. Phys. A: Math. Gen., 23*, 2631–2650.

Treves, A., & Rolls, E. T. (1991). What determines the capacity of autoassociative memories in the brain? *Network, 2*, 371–397.

Treves, A., & Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus, 4*, 374–391.

Tsodyks, M. (2005). Attractor neural networks and spatial maps in hippocampus. *Neuron, 48*(2), 168–169.

Waugh, F. R., Marcus, C. M., & Westervelt, R. M. (1990). Fixed-point attractors in analog neural computation. *Physical Review Letters, 64*(16), 1986–1989.

Wersing, H., Beyn, W. J., & Ritter, H. (2001). Dynamical stability conditions for recurrent neural networks with unsaturating piecewise linear transfer functions. *Neural Computation, 13*(8), 1811–1825.

Wills, T. J., Lever, C., Cacucci, F., Burgess, N., & O'Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science, 308*(3), 873–876.

Yi, Z., Tan, K. K., & Lee, T. H. (2003). Multistability analysis for recurrent neural networks with unsaturating piecewise linear transfer functions. *Neural Computation, 15*(3), 639–662.

Zemel, R. S., & Mozer, M. C. (2001). Localist attractor networks. *Neural Computation, 13*, 1045–1064.

Received July 7, 2009; accepted November 12, 2009.