Dynamics of Memory Representations in Networks with Novelty-Facilitated Synaptic Plasticity

Barak Blumenfeld,[1,2] Son Preminger,[1,2] Dov Sagi,[1] and Misha Tsodyks[1,*]
[1] Department of Neurobiology
The Weizmann Institute of Science
Rehovot 76100
Israel

Summary

The ability to associate some stimuli while differentiating between others is an essential characteristic of biological memory. Theoretical models identify memories as attractors of neural network activity, with learning based on Hebb-like synaptic modifications. Our analysis shows that when network inputs are correlated, this mechanism results in overassociations, even up to several memories "merging" into one. To counteract this tendency, we introduce a learning mechanism that involves novelty-facilitated modifications, accentuating synaptic changes proportionally to the difference between network input and stored memories. This mechanism introduces a dependency of synaptic modifications on previously acquired memories, enabling a wide spectrum of memory associations, ranging from absolute discrimination to complete merging. The model predicts that memory representations should be sensitive to learning order, consistent with recent psychophysical studies of face recognition and electrophysiological experiments on hippocampal place cells. The proposed mechanism is compatible with a recent biological model of novelty-facilitated learning in hippocampal circuitry.

Introduction

Neural network models of associative memory suggest that memories are represented as stable network activity states called attractors (Hopfield, 1982; McNaughton and Morris, 1987; Treves and Rolls, 1992; Amit, 1995; Brunel, 2005). When a stimulus pattern is presented to the system, the network dynamics are drawn toward the attractor that corresponds to the memory associated with that stimulus. The formation of the attractor landscape is achieved by overlapping patterns of synaptic modifications adhering to the Hebbian paradigm (Hebb, 1949) such that each synapse is involved in the storage of multiple related memories. Inevitably, this common synaptic representation implies interactions, or associations, between memories stored in the same network.

Attractor dynamics is commonly believed to underlie the persistent activity observed in the neocortex during working memory experiments (Fuster and Alexander, 1971; Miyashita, 1988; Miyashita and Chang, 1988; Chafee and Goldman-Rakic, 1998; Ericson and Desimone,

1999). It was also suggested that the hippocampal region CA3 could be an anatomical substrate where attractor neural networks (ANN) reside (Treves and Rolls, 1994). Several models of place-specific activity in hippocampus based on ANNs were proposed in the literature (e.g., Tsodyks and Sejnowski, 1995; Samsonovich and McNaughton, 1997), even though experimental support for them is largely indirect. A new evidence for attractor dynamics emerged in a recent study by Wills et al. (2005) where activity of hippocampal place cells was recorded while rats explored a set of environments of different shapes. Initially, when the rats explored a circle- and a square-shaped environment, the similarity between the representations of these two environments by the population activity of hippocampal cells (i.e., the place fields for all the cells that were active in these environments) was low. Later, the rats explored a set of environments of intermediate shapes between the circle and the square. Surprisingly, no gradual change in the representations was observed; instead, most cells abruptly changed their place fields at the same intermediate shape (see Figure 1A). Wills et al. (2005) interpreted their results as evidence for the presence of two distinct attractors corresponding to the circle- and square-shaped environments.

The generality of the Wills et al. (2005) observations was challenged by the study of Leutgeb et al. (2005), in which similar experiments yielded very different results. 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 (see Figure 1B), with different cells changing their place field properties at different paces. It was also observed that the resulting representations were significantly skewed toward the first shape in the sequence, such that the overlap between the circular and square representations was increased significantly compared to the pretraining phase, indicating their partial merging (Figure 1B).

An important difference between the ways the experiments were performed in these two studies concerned the order in which the animals explored the morphed environments. While in the Leutgeb et al. (2005) study the exploration was done in consecutive order, from circle to square or back, in the study by Wills et al. (2005), a scrambled-order exploration was performed. Taken together, these two studies indicate that place representations can be modified due to exposure to morphed environments, and that the changes in representation are contingent upon a particular order of exposure.

Our recent psychophysical study (Preminger et al., 2005; S.P., D.S., and M.T., unpublished data) indicates that similar processes can occur in a very different system and can have a direct and dramatic effect on behavior. In this experiment subjects repeated an identification task on a sequence of morphed faces, gradually transforming from one prelearned face ("source") to another initially distinguishable face ("target," see Figures 2A and 2B). This resulted, under certain conditions, in

*Correspondence: misha@weizmann.ac.il
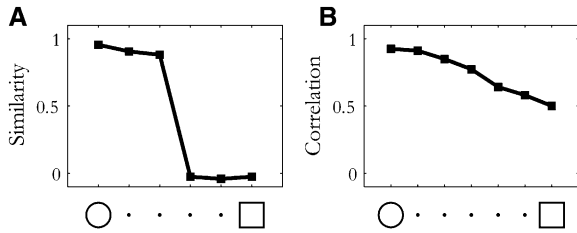[2] These authors contributed equally to this work.

Figure 1. Representations of Morph Environments in Hippocampal Cells

(A) Similarity between hippocampal representations of a sequence of morph environments and the representation of the first environment in the sequence. (Adapted from Figure 3A of Wills et al., 2005). (B) Correlation between the hippocampal representations of a sequence of morph environments and the representation of the first environment in the sequence. (Adapted from Figure 4E of Leutgeb et al., 2005).

partial merging: an increasing fraction of the morph faces were identified as the source face (Figure 2C). Importantly, this effect only occurred when the faces were presented to subjects in a gradual order, albeit interrupted by numerous other familiar and unfamiliar faces. Presentation of the same sequence of images in a random order yielded no significant change in identification (Figure 2C), again indicating the important role of learning order.

In the current contribution, we show that ANN models allow memory representations of correlated stimuli that depend on learning order. To this end, we introduce a network that stores incoming patterns via novelty-

facilitated modifications of synaptic connections. When a quasicontinuous set of patterns (morph sequence) is stored in the network, the resulting attractor representations strongly depend on the order of exposure. In particular, we show that when morph patterns are presented in a gradually increasing order, the network exhibits a stronger tendency for merged representations compared with random order. These results provide a unified theoretical explanation for the experimental observations described above and provide a general framework for studying the effects of context on memory representation.

Results

Throughout the analysis, we use the Hopfield network (Hopfield, 1982) as our modeling framework for studying attractor dynamics of long-term memory. We describe the neurons in the network with binary variables, $S_i(t)$ (+1 if the i-th neuron is active at time t, and −1 otherwise). At every time step, the state of each neuron is updated according to the total synaptic input it receives from the rest of the network via synaptic connections with the strengths J_{ij} (for a connection from neuron number j to neuron number i):

$$S_i(t+1) = \text{sign}\left(\sum_{j=1}^{N} J_{ij} S_j(t) \right) \quad (1)$$

where sign($\cdot$) is the sign function sign(x) = 1 if $x > 0$, sign(x) = −1 if $x < 0$. We assume that stimuli to be stored
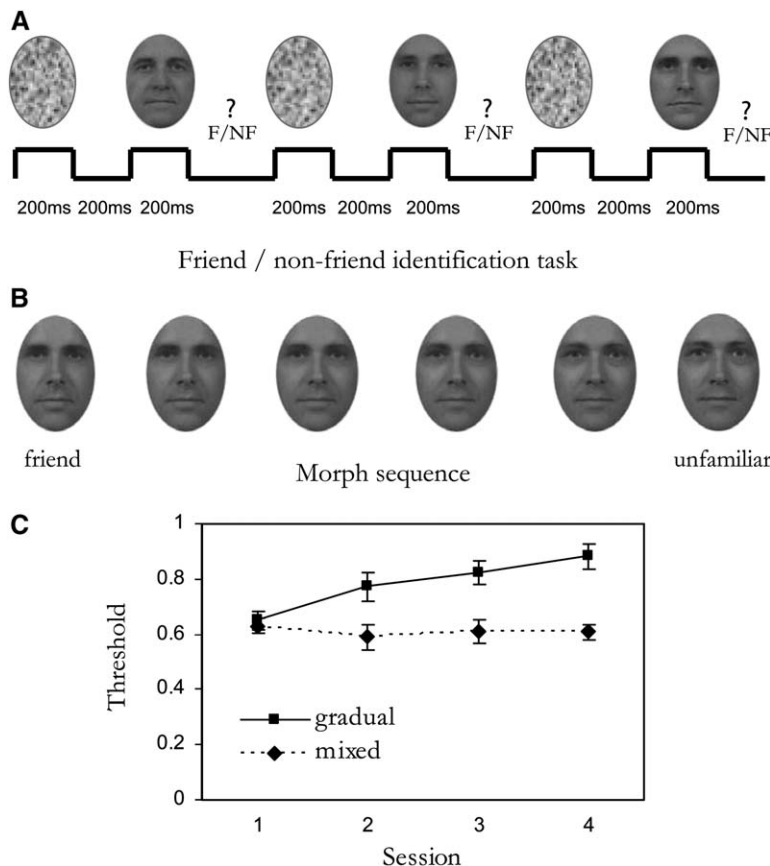

Figure 2. Memory and Identification of Morphed Faces: Gradual versus Mixed Learning Order

In our psychophysical experiments (Preminger et al., 2005; S.P., D.S., and M.T., unpublished data), human subjects were first trained with feedback to identify a set of four faces as friends versus other nonfriend faces. Then, repeatedly in daily sessions, subjects continued to perform this friend/nonfriend identification task without feedback (A). During this practice, one of the friend faces gradually transformed, via 50 intermediate faces, to an initially unfamiliar face throughout each session (morph sequence, B). Thus, at every session each subject saw the full morph sequence in an increasing gradual order, interleaved with other prelearned and nonprelearned faces. Another group of subjects performed almost exactly the same protocol, differing in that during identification sessions, the morph sequence was presented in random order. (C) Results of subjects who performed the gradual protocol are depicted by the solid line (n = 7); results of the mixed protocol group are depicted by the dashed line (n = 4). For each group and each one of four training sessions, the average identification threshold (i.e., portion of the morph sequence identified as friend) across subjects is presented. Initial identification thresholds of subjects shown in this graph range between 0.54 and 0.78. Error bars = SE.

in memory are encoded as network activity patterns, also binary, denoted as ξ_i^μ (μ being the index of the corresponding pattern). In the original Hopfield model, the patterns are stored (or learned) by adding Hebb-like terms to the strength of each synapse such that when all of the patterns are learned, the connections acquire the following values:

$$J_{ij} = \frac{1}{N} \sum_\mu \xi_i^\mu \xi_j^\mu \qquad (2)$$

The memories are retrieved by providing an initializing input and then allowing the network to evolve autonomously by updating the neuron's activity states according to Equation 1, until they converge to one of the stable states, or attractors. Indeed, it was shown in earlier work that all the memories stored in the network can be reliably retrieved provided that the patterns representing them are uncorrelated and that the total number of patterns is smaller than the critical capacity (Hopfield, 1982; Amit et al., 1985b).

We will use the Hopfield formulation and its extension (described below) 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 3). The sequence is constructed by first choosing the source (ξ^0) and the target (ξ^1) patterns randomly, such that for approximately half of the neurons, the source and the target states match each other ($\xi_i^0 = \xi_i^1$), and for the other half of the neurons, the source and target do not match ($\xi_i^0 = -\xi_i^1$). At every step of morphing, we gradually update the state of a fixed number of randomly chosen neurons, for which there is a remaining mismatch between the previous state and the target one, such that the target pattern is obtained at the end.

Storing a Morph Sequence in a Standard Hopfield Network Yields a Single Attractor

The standard analysis identifying the entire set of stored patterns with attractors of the dynamics crucially depends on the assumption that the patterns are uncorrelated, i.e.:

$$\frac{1}{N} \sum_i \xi_i^\mu \xi_i^\nu \approx 0$$

for every $\mu \neq \nu$, such that interference between different patterns during the memory retrieval is minimal. However, for the morph sequence construction described above, the correlation between a pair of patterns (ξ^μ and ξ^ν) decreases gradually from 1 to 0 as the distance between them increases:

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

This results in a much stronger interference between the patterns, and it consequently results in very different behavior of the network when stored memories are being retrieved (see Experimental Procedures A for the mathematical solution of the model).
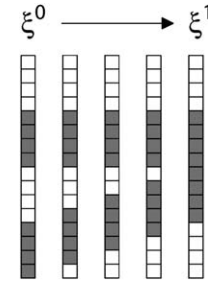

Figure 3. Example of a Morph Sequence with 16 Neurons and Five Patterns

At each pattern, half of the neurons are +1 (shaded boxes) and half are −1 (hollow boxes). Two neurons change sign at each step of the morph.

If we limit ourselves to the initial states of the network that are positively correlated with the morph patterns, our analysis shows that after one step of the dynamics (we consider parallel dynamics in which all of the neurons are updated simultaneously), the network activity states pass exclusively through the morph sequence. The entire dynamics can therefore be described as a sequence of morph patterns, with indices μ_0, μ_1, μ_2…. Eventually the dynamics converge to a stable steady state, ξ^{μ^*}, which is an attractor of the network. As shown in Experimental Procedures A, in the limit of a large number of patterns, the index μ^* can be found by solving the following equation:

$$\int_0^{\mu^*} d\nu(1 - |\nu - \mu^*|) = \int_{\mu^*}^1 d\nu(1 - |\nu - \mu^*|) \qquad (3)$$

Since the expression under the integral is the correlation of the attractor (μ^*) with other patterns (ν), this condition arises from the requirement that opposing interferences from both sides of the attractor balance out to produce an equilibrium state. It is thus clear that the only solution of this equation is $\mu^* = 0.5$. We conclude that a Hopfield network that stores a sequence of morph patterns possesses a single attractor state that coincides with the middle point in the sequence, $\xi^{0.5}$. This analysis illustrates the emergence of a single attractor representation of an entire sequence of correlated patterns that are stored in a Hopfield network via the standard Hebb-like connectivity rule.

Scaling of Hebbian Terms Allows Multiple Attractors

The experimental results described in the introduction (Leutgeb et al., 2005, and Wills et al., 2005; Preminger et al., 2005) indicate that the attractor representation depends on the training protocol. The Hopfield model, however, cannot account for this effect, since it predicts that morph patterns invariably merge into one attractor representation. We therefore introduce an extended model where patterns are stored into the network with variable weights:

$$J_{ij} = \frac{1}{N} \sum_\mu w_\mu \xi_i^\mu \xi_j^\mu \qquad (4)$$

In this formulation, the Hebb-like terms $\xi_i^\mu \xi_j^\mu$, which encode the memorized patterns, are modulated by

corresponding weights w_μ. The idea behind this approach is that depending on the way the patterns are presented for learning (temporal order, frequency of presentation, temporal proximity, presence of intervening patterns, etc.), they may acquire different "saliencies" that are captured by the weights w_μ in Equation 4. Before delving into the learning process during which the weights are acquired, we first address the issue of how a given distribution of weights can give rise to multiple attractor states.

Similarly to the Hopfield model, the dynamics of network states pass exclusively through patterns of the morph sequence. The condition for the fixed point value of the index μ^*, which generalizes Equation 3, is now:

$$\int_0^{\mu^*} d\nu \ w_\nu(1 - |\nu - \mu^*|) = \int_{\mu^*}^1 d\nu \ w_\nu(1 - |\nu - \mu^*|) \qquad (5)$$

which again expresses the balance of interferences between the attractor and other patterns. Depending on the particular distribution of the weights, this equation may have more than one solution. The attractor states of the network correspond to those solutions that are stable, i.e., the ones to which the network converges via its intrinsic dynamics (Equation 1). For some specific choices of weight distributions, Equation 5 can be solved explicitly and the exact attractor landscape can be obtained. For the uniform weight distribution ($w_\mu \equiv w$ for all μ), Equation 5 reduces to Equation 3, with a single solution ($\mu^* = 0.5$). If however the weights are biased toward the source and the target of the morph sequence ($\mu = 0$ and $\mu = 1$, respectively), the network can acquire two attractors. For instance, if $w_\mu = (\mu - 0.5)^2$, then there are three solutions to Equation 5: two stable ones (attractors) with $\mu^* = 0.5 - 1/\sqrt{8}$ and $\mu^* = 0.5 + 1/\sqrt{8}$, and an unstable one with $\mu^* = 0.5$. The first attractor represents the first half of the morph sequence (patterns with $\mu < 0.5$), with the second attractor representing the remaining half. This result demonstrates that different distributions of saliency weights induce different attractor landscapes.

Novelty-Facilitated Synaptic Modifications Induce History-Dependent Learning

In order to relate the above analysis to the experimental results on history-dependent memory representations, we propose a particular learning process that demonstrates how saliency weights can be acquired throughout learning. Clearly, a required learning mechanism must depend not only on the current input but also on previous experience. Previous experience can in turn be encoded by the attractor landscape itself at the time of the new exposure. We suggest that this dependency can be captured by saliency weights assigned to different patterns presented for learning (see Equation 4). In particular, we propose that the changes in the weights are determined by the perceived "novelty" of the presented pattern. Intuitively, we think of the pattern's novelty as being a measure of its similarity to an existing attractor, i.e., the pattern is perceived as "novel" if the network does not already possess a similar attractor state, and "not novel," or "familiar," if it is similar to an already existing attractor. Mathematically, one could use the number of mismatches between the presented pattern and the attractor state to which it converges (called the Hamming distance) as a novelty signal. Alternatively, one can use the Hamming distance between the presented pattern and the one obtained after one step of iteration (both rules result in very similar behaviors). The second choice seems to be more biologically realistic, as it does not require the initial state of the network to be remembered until it converges to an attractor. We therefore define the change in the synaptic connection between neurons i and j, after the presentation of a pattern μ, to be proportional to a local Hebb-like term, $\xi_i^\mu \xi_j^\mu$, modulated by the global novelty signal defined above. Importantly, the novelty signal is the same for all the connections in the network. This synaptic modification rule is equivalent to the following update of the saliency weight, w_μ, corresponding to the currently presented pattern (see Equation 4):

$$w_\mu \rightarrow w_\mu + \eta H \qquad (6)$$

where H is the Hamming distance between ξ^μ, the pattern presented for learning, and the state of the network after one step of the dynamics. η is a positive constant that defines the speed of learning.

With this learning algorithm, we can simulate the effect of an arbitrary learning protocol on the formation of the attractor states of the network. One can already see from this formulation that the novelty-based learning acts to counterbalance the merging of morph patterns into a single representation. Indeed, if we consider the case where patterns are stored in the network with uniform saliency weights, then, as demonstrated above, the network possesses a single attractor at the middle of the morph sequence. If the network is subsequently exposed to the same sequence, the novelty factor will be higher for the patterns that are farther away from the attractor, so that the weights will become more biased toward the ends of the sequence. As we saw above, this bias may split the attractor in two. In fact, we expect that repeated exposure to the morph patterns will eventually result in a formation of multiple attractor states covering the morph sequence.

Learning Dynamics and Memory Representation Depend on the Order of Stimuli Presentation

The experimental results have demonstrated that depending on the protocol, representations of morph patterns can either "converge" toward each other or keep a categorical distinction between patterns. Here we show how these findings can be captured by the extended Hopfield model combined with the novelty-dependent learning mechanism. To this end, we examine the performance of our model by implementing learning protocols similar to the ones used in the above mentioned studies of hippocampal spatial representations (Leutgeb et al., 2005; Wills et al., 2005). Specifically, we simulate the learning of morph sequence patterns in our attractor network model with novelty-based learning by presenting morphed patterns according to one of two protocols: gradual protocol, where the morph sequence is presented in a gradually increasing index order, or mixed protocol, where the morph patterns are presented in a randomly ordered sequence. The simulation results
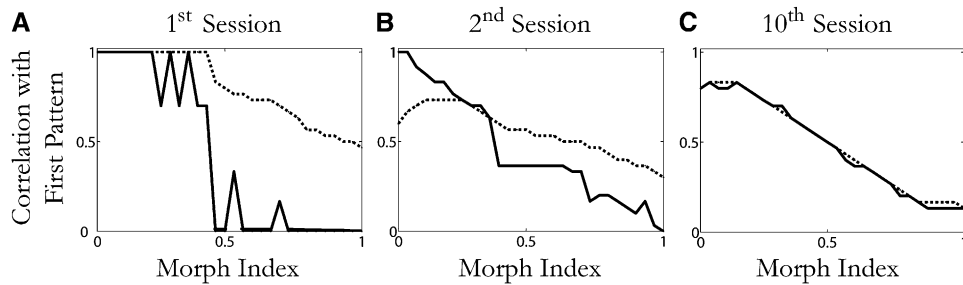
Figure 4. Simulation of Novelty-Based Learning of Morph Patterns in Gradual and Mixed Presentation Order

Before learning, the source and target patterns had equal saliency weights (=1); other patterns had zero weights. In every learning session, a set of 30 morph sequence patterns was presented in either an increasing gradual order (dashed line) or mixed order (solid line; every session, a different random order of the morph patterns was used). After each pattern was presented, its attractor representation was obtained using Equation 1, and the corresponding weight was updated according to Equation 6. The learning rate, η, was 0.5 (the Hamming distance was normalized such that the distance between the source and target patterns is 1). The plots show the correlation between the source pattern and the attractor representation of all patterns in the first (A), second (B), and tenth (C) learning sessions.

are shown in Figure 4. At the beginning of the learning procedure, two random uncorrelated patterns (source and target, corresponding to circle- and square-shaped environments) were stored in the network with equal initial weights, to represent the prelearned circle and square environments. Next, the morphed patterns were presented according to the gradual or mixed protocols. Learning was performed after presentation of each pattern, based on Equation 6. As seen in Figure 4A, after one learning session of the gradual protocol, where each of the morph patterns was presented once, the representation of the last morph pattern exhibited a marked positive correlation with the representation of the first pattern, similarly to the experimental findings of Leutgeb et al. (2005) shown in Figure 1B. On the other hand, a simulation of the mixed protocol produced a categorically-shaped representation where groups of patterns on both edges of the sequence were merged into a single attractor and attractors of the first and last patterns were the same as before the training, corresponding to the findings of Wills et al. (2005) shown in Figure 1A. To investigate the behavior of the learning rule with repetition of the learning, we continued to simulate the protocols, repeating the gradual or mixed presentation correspondingly for multiple sessions. We observed that for both protocols, as the presentation of the patterns was repeated, each pattern was more likely to be drawn to itself, although in different paces, and the correlation between the representations of the source and target decreased (Figures 4B and 4C). This is similar to the findings of Leutgeb et al. (2005), who observed a decrease in correlation between the representations of the two end shapes after multiple repetitions of the gradual protocol.

The simulations presented above indicate that gradual-order presentation of correlated patterns results in stronger merging effects than random-order exposure to the same patterns. This conclusion is in broad agreement with the experimental observations on hippocampal place representation and face recognition. In the next two sections we present a more rigorous analysis in order to ascertain that the difference between gradual-order and random-order exposure is a general feature of the model and does not depend on the fine details of the learning procedures.

Relationship Between the Saliency Factors and Attractors in the Network

In order to understand the dynamics of representations throughout the learning process, we need to establish a theory that allows an estimation of the number of attractors and their locations for a general-form distribution of the saliency weights, w_μ. A central notion in our analysis is the energy function (Hopfield, 1982) that depends on the position of the current activity state of the network on the morph sequence μ (see Experimental Procedures B):

$$E(\mu) = -\frac{1}{2}\int_0^1 d\nu \; w_\nu (1 - |\mu - \nu|)^2 \qquad (7)$$

Importantly, when the network state evolves in time according to the update equations (Equation 1), the energy function cannot increase, so in order to determine the position of attractors, one should only find its local minima. This is illustrated in Figures 5A and 5B where we show the familiar example with constant weights ($w_\mu \equiv 1$). The energy function has a single local minimum at $\mu = 0.5$, as expected from our earlier analysis (see Equation 3). The next example, in Figure 5C, represents a class of weight functions in which the two edges of the morph sequence are accentuated. The energy function (Figure 5D) shows two local minima corresponding to two attractor states, which could represent the experimental results of Wills et al. (2005) (see Figure 4A).

The analysis of the energy function allows us to relate the structure of the attractor landscape to the general shape of w_μ (Experimental Procedures B). We partition the morph sequence into distinct intervals of two types: salient intervals, defined to be intervals where $w_\mu > 0.5$, and nonsalient intervals, where $w_\mu < 0.5$ (for mathematical convenience, we normalize the weights). Our analysis shows that attractors can exist only in the salient intervals and that there is at most one attractor per salient interval. Salient intervals are marked by black line segments in Figure 5 (e.g., the two salient intervals in Figure 5C, which correspond to two attractors in Figure 5D). In general, not every salient interval has an attractor. This is illustrated by the example shown in Figure 5E where the weight function (a 12th order
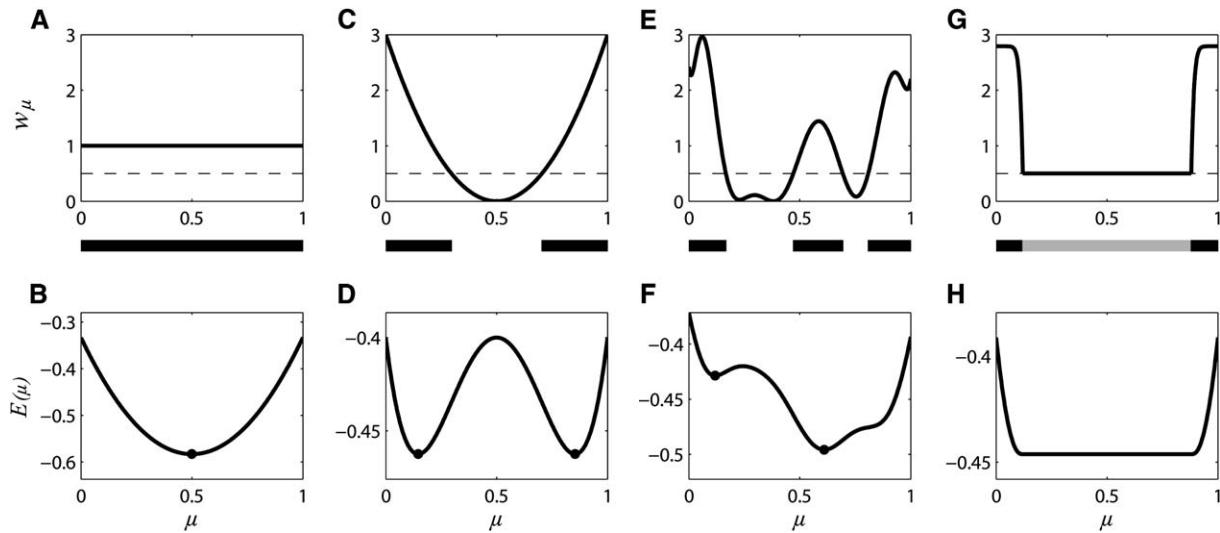
Figure 5. Representative Examples of Weight and Energy Functions

(A), (C), (E), and (G) show the weight functions w_μ. The dashed line is plotted at the critical value of $w_\mu = 0.5$. The black line segments below the panels indicate the salient intervals ($w_\mu > 0.5$), where attractors can occur. The gray line segment below panel (G) indicates a semisalient interval ($w_\mu = 0.5$) where a line attractor can occur. (B), (D), (F), and (H) show the corresponding energy functions, $E(\mu)$. The local minima, identified with the attractors, are indicated by the black dots.

polynomial) has three salient intervals and the energy function has only two local minima (Figure 5F). In Experimental Procedures B we derive a formal condition for the formation of an attractor over a salient interval, which depends on the ratio between the "saliency" of the salient interval and the saliency of other intervals.

Finally, we consider an important example of the weight distribution that is qualitatively different from the previous ones. We choose w_μ to be precisely 0.5 over some interval, which we refer to as a semisalient interval. The weight function in Figure 5G has one semisalient interval at the center, which is indicated by the gray line segment below the figure. Our analysis shows that a semisalient interval can be covered by a continuous set of attractors (see Experimental Procedures B for an exact condition), similar to our simulation results (Figure 4C). This solution is known as a "line attractor" and received much attention in other contexts (e.g., orientation tuning: Ben-Yishai et al., 1995; Blumenfeld et al., 2006; oculomotor integration: Seung, 1996; Bayesian inference: Deneve et al., 1999; see Ermentrout, 1998 for a general discussion of continuous attractors). The energy function at a line attractor is flat, and all points there are marginally stable fixed points (Figure 5H). As shown in Experimental Procedures B, the condition for the existence of a line attractor is always satisfied for a symmetric weight function with a semisalient interval at the center, as in Figure 5G, regardless of the width of the semisalient interval, or the values of the weight function outside the semisalient interval. Thus, it is possible to have an arbitrarily wide line attractor, approaching the entire morph sequence.

Protocol-Dependent Dynamics of Merging and Splitting of Attractors

Equipped with the analytical tools developed above, we can now formulate the following general properties of learning dynamics in our model:

(1) The only way the novelty-based learning can stop is if every pattern is an attractor (see Equation 6). Thus, if the network has unlimited time for learning in a stationary environment, when no new patterns are introduced, the learning dynamics can only converge to a continuous line attractor representation of a morph sequence. Indeed, such a behavior can be observed in Figure 4, both for the mixed and gradual protocols.

(2) If the network is trained for a limited time, the memory representation changes dynamically and at any given time constitutes a mixture of merged and split attractor states. The precise composition of attractors depends not only on the correlation structure between the patterns but also on the precise order in which they are presented for learning.

We observed in simulations (Figure 4) that gradual exposure to morph patterns results in a stronger tendency for merging as opposed to random-order exposure. The generality of this observation can now be confirmed by a rigorous analysis in the case where each of the morph patterns is presented once to a network with no previously stored memories; the attractor landscape at any given time is determined exclusively by learning dynamics.

We can prove that during a gradual, fixed-order exposure, a single attractor representation emerges and is maintained throughout the session (see Experimental Procedures C). This is illustrated in Figure 6A, for which we simulated the gradual protocol and plotted the attractors' position along the sequence after each subsequent presentation. Initially, a single attractor is formed near the source pattern. As the training progresses, this attractor gradually moves toward the target, following an almost-linear path and reaching the final position of $\mu \approx$ 0.7. The (normalized) saliency weights at 33%, 67%, and

Gradual

A

B

C

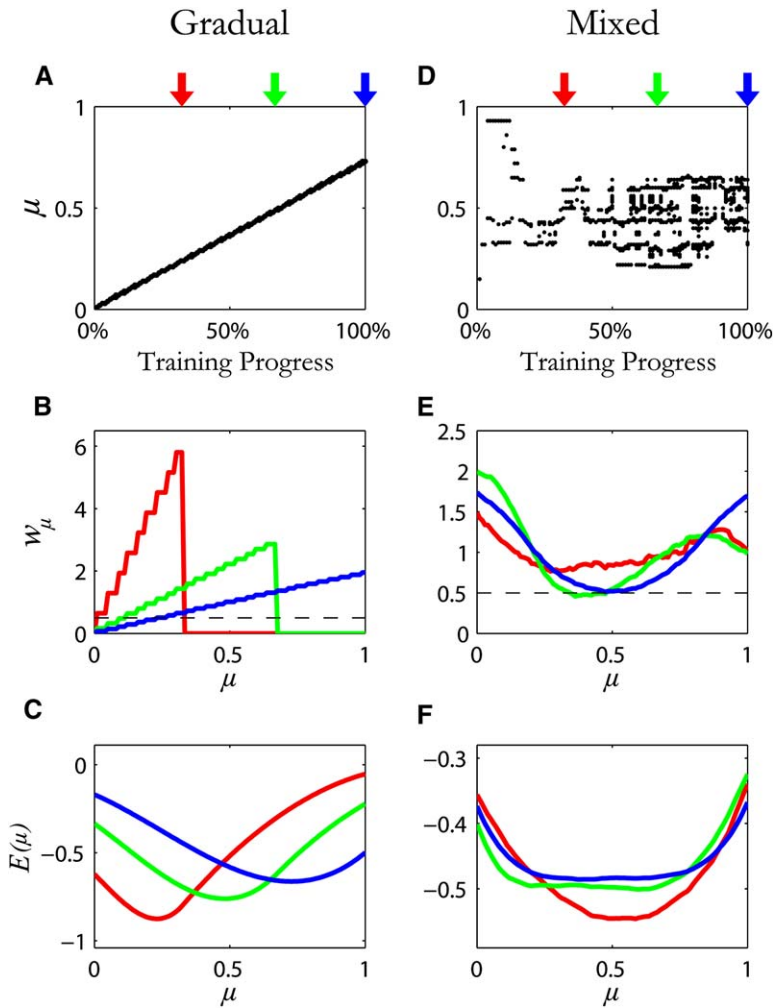## Mixed

D

E

F

Figure 6. Simulations of Learning with the Gradual and Mixed Protocol

Location of attractors (black dots) as a function of training progress in the gradual (A) and mixed (D) protocols. The number of patterns was 100. Weights are shown at 33% (red), 67% (green), and 100% (blue) of the training with gradual (B) and mixed (E) protocols. To allow comparison with the critical value of $w_\mu = 0.5$ (dashed line), the weights were normalized to have an integral of 1. In addition, the weights in (E) were smoothed with a Gaussian kernel ($\sigma = 8$ patterns). Energy, calculated with normalized weights, at 33% (red), 67% (green), and 100% (blue) of the training with gradual (C) and mixed (F) protocols is shown.

100% of the training (Figure 6B) also show an approximately linear structure. The existence of a single attractor can be inferred from the pattern of the saliency weights, since at any given moment there is only one salient interval. Correspondingly, the energy function always has one local minima, which moves in the direction of the target pattern as the training progresses (Figure 6C).

In contrast, when the patterns are presented using the mixed protocol, a symmetric weight distribution characterized by multiple attractor representation emerges (see Figure 6D). The precise position of the attractors and their number changes during training and depends on the specific realization of the random order of the presentation. However, even at a single typical training session, the weights evolve in a specific manner, which is revealed by plotting a smoothed version of the weights at different points of the training (Figure 6E). It can be seen that the weights approach a symmetric shape, of which the central interval is a semisalient interval. According to our previous analysis, such a shape supports a line attractor (Figures 5G and 5H). This trend is even more apparent when the energy function is considered (Figure 6F); note that the energy gets flatter as the training progresses. The difference between the gradual and mixed protocols is clearly seen from comparing Figures

6C and 6F: the gradual protocol tends to shift the minimum of the energy, therefore maintaining a merged representation; the mixed protocol tends to flatten the energy, which results in splitting the attractor representation.

Discussion

The model proposed in this contribution puts forward an analytical framework for studying network mechanisms of memory association, grouping, and context, and how they can dramatically affect the resulting memory representations. These concepts can be subsequently implemented with more biologically detailed models. Indeed, the computational principles of the Hopfield network have been shown to be applicable in more biologically realistic architectures (e.g., 0/1 neurons: Tsodyks and Feigel'man, 1988; firing rate neurons: Treves, 1990; spiking neurons: Amit and Brunel, 1997; see Brunel, 2005 for a review).

Our model shows how selective merging and unmerging of different items (such as sensory objects, spatial environments, etc.) could occur in a long-term memory system. The analysis demonstrates that the merging dynamics should necessarily occur whenever the stored

memories are represented by the neural system as correlated patterns of neuronal activation. We suggest that this tendency for merging is counterbalanced by a novelty-based learning mechanism. As a result, the memory representation changes dynamically and at any given time constitutes a mixture of merged and split attractor states. The precise composition of attractors depends not only on the correlation structure between the patterns but also on the precise order in which they are presented for learning. In particular, gradual exposure to morph patterns causes a stronger tendency for merging as opposed to random-order exposure, in accordance with the experimental observations on hippocampal place representations. We also described our psychophysical results from morphed faces recognition experiments, indicating that similar learning dynamics occur in the visual system. Further work is required to combine the presented analysis with modeling of the recognition process in order to develop a comprehensive theoretical description of these experiments.

The model that we propose is based on two basic underpinnings. First, memories are stored via a Hebb-like process of long-term synaptic modifications and retrieved via internal dynamics of patterned network activity. This theoretical idea is supported by a vast experimental literature on both activity-dependent synaptic plasticity (Bliss and Lomo, 1973; Bear and Malenka, 1994) and persistent activity in the neocortex (Fuster and Alexander, 1971; Miyashita, 1988; Miyashita and Chang, 1988; Chafee and Goldman-Rakic, 1998; Ericson and Desimone, 1999). Persistent activity in the cortex is generally believed to result from attractor dynamics, although recent experiments unveiled the possible role of intrinsic neuronal properties (Fransen et al., 2006). The evidence for attractor dynamics underlying spatial representations in hippocampus is not as direct. While Wills et al. (2005) interpreted their results as supporting the attractors theory, the different results of Leutgeb et al. (2005) raise the question of how general this interpretation is. One of the contributions of our study is the reconciliation between these studies by demonstrating the flexibility of the attractor representations and their high sensitivity to learning history.

The second quiddity of our model is that the process of encoding items into memory is strongly affected by their perceived novelty (Tulving et al., 1996), and furthermore, that this effect is mediated by a global novelty signal that determines the strength of synaptic modifications. Experimental work has shown that novel stimuli evoke characteristic responses in a widespread network of brain regions (see Friedman et al., 2001; Ranganath and Rainer, 2003; and Nyberg, 2005, for reviews). Novelty-induced activation of the hippocampus, related medial-temporal lobe regions, and the prefrontal cortex has been associated with the enhanced encoding of memory items (Kirchhoff et al., 2000; Meltzer and Constable, 2005). In addition, it was shown that neuromodulatory systems such as the dopaminergic and cholinergic systems are activated by novel stimuli (Schultz, 2002), resulting in the release of a neuromodulator in numerous brain regions, which in turn can facilitate synaptic plasticity (Gu, 2002; Li et al., 2003; Otani et al., 2003). The proposal that a mismatch between external input and internal representations drives learning

has been considered in the computational literature (e.g., Carpenter and Grossberg, 2003). Similar ideas are used in contrastive Hebbian learning (Movellan, 1990) and contrastive divergence learning (Hinton, 2002), where no global novelty signal is introduced; instead, Hebbian and anti-Hebbian modifications are applied for subsequent phases of neural dynamics. Novelty-induced learning was previously implemented in hippocampus modeling by Hasselmo and Schnell (1994) and Hasselmo et al. (1995), where an acetylcholine-mediated novelty signal was estimated based on the global activity level upon presentation of an input.

In our model, novelty-facilitated learning is implemented as a scaling of Hebbian modifications by the global signal calculated as a difference between the input and the corresponding activity pattern produced by network dynamics. This mathematical rule is consistent with recently introduced biologically inspired framework that suggests a specific functional loop between the hippocampus and the VTA, which acts to facilitate the encoding of novel stimuli (Lisman and Grace, 2005). In particular, Lisman and Grace proposed that the novelty signal is initiated in the CA1 field of the hippocampus by computing the mismatch between sensory representation of a stimulus, received via the direct connection from the entorhinal cortex, and its internal representation computed in the dentate gyurs/CA3 regions and conveyed via the CA3/CA1 connection. This step is analogous to the computation of the distance between the input and its processed network representation in our model. The novelty signal in the Lisman and Grace model is next conveyed via a polysynaptic pathway to the VTA, where it evokes firing of dopaminergic neurons. These neurons project back to the hippocampus, where the release of dopamine facilitates LTP and consequently enhances the further encoding of the stimulus. Analogously, in our model, the synaptic weights are updated proportionally to the novelty signal (Equation 6).

The analysis of our model allowed us to obtain some general predictions about the effects of the overall context and learning protocol on the resulting memory dynamics. In particular, we predict that given infinite exposure time and a stable sensory environment, the system could find a way to store the memory items in the most faithful way possible, i.e., it would create a stable memory state (network attractor) for each item. Both of these requirements are of course not realistic for a biological system that has to operate under the conditions of limited time available for learning and frequent arrival of new inputs. One can therefore interpret the learning in our model as a continuous process of adaptation of the memory system to the ever-changing sensory environment. We believe that our model, being accessible to quantitative analysis, provides a convenient framework for studying the effects of context on long-term memory and sensory perception. The model can be used as a guiding tool for designing new memory experiments with different learning protocols. In addition to the novelty factor, one could consider other global signals that affect the memory, such as the ones originating in other sensory modalities, or the feedback signals that carry the evaluation of the system's performance on the memory task.

Experimental Procedures

A. Analysis of Dynamics

In this section we analyze the dynamics of the Hopfield model in which a morph sequence is stored and derive the fixed point equation. We will consider the extended model with weights (Equation 4). The original Hopfield model (Equation 2) arises as a particular case for which $w_\mu \equiv 1$.

As usual in the theory of Hopfield networks (Amit et al., 1985a), it is helpful to consider the overlap between the state of the network and the patterns stored in the connectivity matrix. We therefore define:

$$m^\mu(t) = \frac{1}{N}\sum_i \xi_i^\mu S_i(t) \tag{A1}$$

to be the overlap between the network state $S(t)$ and pattern ξ^μ. Using this definition, the state of the network at time $t + 1$ can be written as:

$$S_i(t+1) = sign\left(\sum_\nu \xi_i^\nu w_\nu m^\nu(t)\right)$$

which means that $S(t + 1)$ depends on $S(t)$ only through $m^\mu(t)$. Thus, given $m^\mu(t)$, $m^\mu(t + 1)$ is given by:

$$m^\mu(t+1) = \frac{1}{N}\sum_i sign\left(\sum_\nu \xi_i^\mu \xi_i^\nu w_\nu m^\nu(t)\right) \tag{A2}$$

Our analysis proceeds by simplifying the expression on the right-hand side of Equation A2. We define:

$$I_i^\mu = \sum_\nu \xi_i^\mu \xi_i^\nu w_\nu m^\nu(t)$$

and divide the summation over the neuron indexes i into three parts

$$m^\mu(t+1) = \frac{1}{N}\sum_{i\in F} sign(I_i^\mu) + \frac{1}{N}\sum_{i\in C_0^\mu} sign(I_i^\mu) + \frac{1}{N}\sum_{i\in C_1^\mu} sign(I_i^\mu) \tag{A3}$$

with the sets of indexes F, C_0^μ, and C_1^μ defined below. The set F is defined as the set of indexes i for which the source and target match ($\xi_i^0 = \xi_i^1$). For $i\in F$:

$$I_i^\mu = \sum_\nu w_\nu m^\nu(t)$$

and under the assumption $m^\nu(t)$, $w_\nu > 0$ for all ν, it follows immediately that:

$$\frac{1}{N}\sum_{i\in F} sign(I_i^\mu) = \frac{1}{2} \tag{A4}$$

Next, we consider neurons for which the source and target do not match ($\xi_i^1 = -\xi_i^0$). For each such neuron we assign an index, $0 < \rho_i \le 1$, which denotes the location in the morph sequence in which the neuron changes sign (i.e., ρ_i is the smallest index μ for which $\xi_i^\mu = -\xi_i^0$). We define the set C_0^μ to be the set of neuron indexes for which $\mu < \rho_i$, and C_1^μ to be the set of indexes for which $\mu \ge \rho_i$. For $i\in C_0^\mu$, it holds that:

$$I_i^\mu = \sum_{\nu<\rho_i} w_\nu m^\nu(t) - \sum_{\nu\ge\rho_i} w_\nu m^\nu(t)$$

and therefore:

$$\frac{1}{N}\sum_{i\in C_0^\mu} sign(I_i^\mu) = \frac{1}{N}\sum_{i\in C_0^\mu} sign\left(\sum_{\nu<\rho_i} w_\nu m^\nu(t) - \sum_{\nu\ge\rho_i} w_\nu m^\nu(t)\right) \tag{A5}$$

The expression inside the sign function depends on the set of neuron indexes i only through ρ_i. We can therefore replace the summation over $i\in C_0^\mu$ with summation over the corresponding values of ρ_i. Since at every morph step we update $N/(2P)$ neurons (P is the number of patterns), the summation takes the form:

$$\frac{1}{N}\sum_{i\in C_0^\mu} sign(I_i^\mu) = \frac{1}{2P}\sum_{\rho_i>\mu} sign\left(\sum_{\nu<\rho_i} w_\nu m^\nu(t) - \sum_{\nu\ge\rho_i} w_\nu m^\nu(t)\right) \tag{A6}$$

Using the assumption of large P, and the fact that ρ_i is a uniformly distributed random variable, we replace sums with integrals to get:

$$\frac{1}{N}\sum_{i\in C_0^\mu} sign(I_i^\mu) = \frac{1}{2}\int_\mu^1 d\rho\, sign\left(\int_0^\rho d\nu w_\nu m^\nu(t) - \int_\rho^1 d\nu w_\nu m^\nu(t)\right) \tag{A7}$$

Under the assumptions $m^\nu(t)$, $w_\nu > 0$, the expression:

$$\int_0^\rho d\nu w_\nu m^\nu(t) - \int_\rho^1 d\nu w_\nu m^\nu(t)$$

is a monotonically increasing function of ρ, which crosses zero at a unique point $\rho = \mu_t$ that satisfies:

$$\int_0^{\mu_t} d\nu w_\nu m^\nu(t) = \int_{\mu_t}^1 d\nu w_\nu m^\nu(t) \tag{A8}$$

From these observations, it follows that Equation A7 can be rewritten as:

$$\frac{1}{N}\sum_{i\in C_0^\mu} sign(I_i^\mu) = \frac{1}{2}\int_\mu^1 d\rho sign(\rho - \mu_t) \tag{A9}$$

Similarly, we can calculate the sum over the remaining set of indexes, C_1^μ, to get:

$$\frac{1}{N}\sum_{i\in C_1^\mu} sign(I_i^\mu) = -\frac{1}{2}\int_0^\mu d\rho sign(\rho - \mu_t). \tag{A10}$$

We are now ready to sum Equations A4, A9, and A10, and by substituting the sum in Equation A3, we get:

$$m^\mu(t+1) = \frac{1}{2} + \frac{1}{2}\int_\mu^1 d\rho\, sign(\rho - \mu_t) - \frac{1}{2}\int_0^\mu d\rho\, sign(\rho - \mu_t) \tag{A11}$$

Performing the integrals, we obtain:

$$m^\mu(t+1) = 1 - |\mu - \mu_t| \tag{A12}$$

where μ_t is the solution of Equation A8. If we now assume $m^\nu(0)$, $w_\nu > 0$, and substitute $t = 0$, $\mu = \mu_0$ in Equation A12, we find that $m^{\mu_0}(1) = 1$. From the definition of m^μ (Equation A1), it follows that $m^\mu = 1$ if and only if the state of the network is ξ^μ. Thus, we conclude that after a single iteration of the network dynamics, the state of the network is $S(1) = \xi^{\mu_0}$. Inductively, this implies that the entire dynamics of the network $S(1)$, $S(2)$, $S(3)$,… can be described by the sequence ξ^{μ_0}, ξ^{μ_1}, ξ^{μ_2},…., such that:

$$\int_0^{\mu_{t+1}} d\nu w_\nu(1 - |\nu - \mu_t|) = \int_{\mu_{t+1}}^1 d\nu w_\nu(1 - |\nu - \mu_t|) \tag{A13}$$

which follows from Equation A8 and Equation A12. Attractor states of the network are fixed points of Equation A13, which is Equation 5 of the results.

B. Analysis of Energy

In this section we derive and analyze the energy function for the Hopfield model in which a morph sequence is stored. In addition we show how the energy can be used to estimate the number of attractors and their locations.

Properties of the Energy Function

We use the energy function introduced by Hopfield (1982):

$$E(S) = -\frac{1}{2NP}\sum_{i,j} J_{ij} S_i S_j \tag{B1}$$

Our analysis (Experimental Procedures A) has shown that the network states are in fact morph sequence patterns. Therefore, the energy can be written as a function of the location of the network state in the morph sequence, which we denote by μ. By exploiting the structure of the connectivity matrix (Equation 4), we find that in the limit of large P, Equation B1 can be written as:

$$E(\mu) = -\frac{1}{2} \int_0^1 d\nu \; w_\nu (1 - |\mu - \nu|)^2 \qquad \text{(B2)}$$

The dynamics of the network $\mu_1, \mu_2,\dots$ are related to the energy through the following properties:

Property (1): μ_t changes against the gradient of the energy function (i.e., if $E'(\mu_t) > 0$ then $\mu_{t+1} < \mu_t$; if $E'(\mu_t) < 0$ then $\mu_{t+1} > \mu_t$).

Property (2): μ_t converges to the nearest local minima of the energy according to Property (1) (i.e., if $E'(\mu_{t_0}) > 0$, then $\mu_{t_0}, \mu_{t_0+1},\dots$ converges to $max\{\mu | \mu \in M, \mu < \mu_{t_0}\}$, where M denotes the set of local minima of the energy; if $E'(\mu_t) < 0$, then $\mu_{t_0}, \mu_{t_0+1},\dots$ converges to $min\{\mu | \mu \in M, \mu > \mu_{t_0}\}$).

For the following analysis it is convenient to define a two-variable function:

$$f(x_0, x_1) = \int_0^{x_1} d\nu \; w_\nu (1 - |\nu - x_0|) - \int_{x_1}^1 d\nu \; w_\nu (1 - |\nu - x_0|) \qquad \text{(B3)}$$

To show Property (1), we note that $f(\mu_t, \mu_{t+1}) = 0$ (see Equation A13), and that $f(\mu,\mu) = E'(\mu)$ (see Equation B2). Thus, the equality:

$$\int_{\mu_t}^{\mu_{t+1}} \frac{\partial}{\partial x_1} f(\mu_t, x_1) dx_1 = f(\mu_t, \mu_{t+1}) - f(\mu_t, \mu_t)$$

Implies:

$$E'(\mu_t) = -2 \int_{\mu_t}^{\mu_{t+1}} d\nu \; w_\nu (1 - |\nu - \mu_t|) \qquad \text{(B4)}$$

from which Property (1) immediately follows.

To show Property (2), we will denote by g the function which maps μ_t to μ_{t+1} [i.e., $\mu_{t+1} = g(\mu_t)$]. Note that g can be defined through f: for any fixed x_0, the function $f(x_0, x_1)$ is a strictly increasing function of x_1, with $f(x_0, 0) < 0$ and $f(x_0, 1) > 0$. Since f is continuous, for any x_0 there is a unique x_1 such that $f(x_0, x_1) = 0$. The function g is therefore the unique function that satisfies $f(\mu_0, g(\mu_0)) = 0$.

Next, we construct the inverse function g^{-1}. Let $\alpha_0 = g(0)$, and $\alpha_1 = g(1)$ (as we will see, $[\alpha_0, \alpha_1]$ will be the domain of g^{-1}). Claim: $\alpha_0 < \alpha_1$. Proof: let α be the point that satisfies:

$$\int_0^\alpha d\nu \; w_\nu - \int_\alpha^1 d\nu \; w_\nu = 0$$

We observe that $f(0,\alpha)$ can be written as:

$$f(0, \alpha) = \int_0^\alpha d\nu \; w_\nu(\alpha - \nu) - \int_\alpha^1 d\nu \; w_\nu(\alpha - \nu) +$$

$$(1 - \alpha) \left(\int_0^\alpha d\nu \; w_\nu - \int_\alpha^1 d\nu \; w_\nu \right)$$

which implies $f(0,\alpha) > 0$. Since $f(0,\alpha_0) = 0$ and f is a continuous and increasing function of its second parameter, it must hold that $\alpha_0 < \alpha$. A similar argument, using $f(1,\alpha)$ can be made to derive $\alpha_1 > \alpha$. Therefore $\alpha_0 < \alpha_1$.

For any $\alpha_0 < x_1 < \alpha_1$, we observe that $f(0, x_1) > 0$ and $f(1, x_1) < 0$, which together with the fact that:

$$\frac{\partial^2 f}{\partial x_0^2} = \begin{cases} -2w_{x_0} & x_0 < x_1 \\ 2w_{x_0} & x_0 > x_1 \end{cases} \qquad \text{(B5)}$$

implies that for any $\alpha_0 < x_1 < \alpha_1$, there is a unique x_0 satisfying $f(x_0, x_1) = 0$. We thus define function $g^{-1}(x_1)$ for $\alpha_0 < x_1 < \alpha_1$ to be the unique function satisfying $f(g^{-1}(x_1), x_1) = 0$. For the end-points α_0, α_1, we define $g^{-1}(\alpha_0) = 0$ and $g^{-1}(\alpha_1) = 1$.

It is now straightforward to verify that g^{-1} is indeed the inverse of g. Since g is invertible and continuous, it must be monotonic. We have already shown that $\alpha_0 < \alpha_1$, i.e., $g(0) < g(1)$; so, g must be a monotonically increasing function. Property (2) follows directly from g being monotonically increasing.

Estimation of Attractor Location

As the analysis above shows, attractors can be identified with local minima of the energy function. To estimate the number and location of the attractors, we analyze the first and second derivatives of this function:

$$E'(\mu) = \int_0^\mu d\nu \; w_\nu - \int_\mu^1 d\nu \; w_\nu - (\mu - \bar{\nu}) \qquad \text{(B6)}$$

$$E''(\mu) = 2w_\mu - 1 \qquad \text{(B7)}$$

where:

$$\bar{\nu} = \int_0^1 d\nu \; w_\nu \nu$$

and we normalize the weights for convenience, such that:

$$\int_0^1 d\nu \; w_\nu = 1$$

The local minima of the energy are solutions to $E'(\mu) = 0$ (which is equivalent to Equation A12) that also satisfy $E''(\mu) > 0$. To estimate the local minima, we observe that Equation B7 implies that the energy is convex whenever $w_\mu > 0.5$, and concave when $w_\mu < 0.5$. Obviously, isolated local minima can occur only at intervals where the energy is convex. Thus, we partition the interval [0,1] into alternating intervals with $w_\mu > 0.5$ (salient intervals), $w_\mu < 0.5$ (nonsalient intervals), and intervals in which $w_\mu = 0.5$ (semisalient intervals).

Using these definitions, the following observations can be made.

(A) There is at least one attractor.
(B) There are no attractors over nonsalient intervals.
(C) Every salient interval has at most one attractor.
(D) A semisalient interval has either no attractors or attractors at all of its points.

Property (A) can be inferred from Equation B6 by verifying that $E'(0) < 0$ and $E'(1) > 0$, and therefore, $E'(\mu)$ crosses zero with a positive slope at some point μ. Properties (B), (C), and (D) follow directly from the convexity properties of the energy at the salient/nonsalient intervals (Equation B7). The precise conditions for attractor formation are given below.

Condition for Formation of an Attractor in a Salient Interval

Let $\varphi_0 < \varphi_1$ denote the edges of the salient interval. For an attractor to be formed in the salient interval, the derivative of the energy, $E'(\mu)$, has to vanish at some point in the open interval (φ_0, φ_1). Since $E'(\mu)$ is strictly increasing on the interval, we can require that $E'(\varphi_0) < 0$ and $E'(\varphi_1) > 0$. In turn, this can be replaced by the condition:

$$\frac{E'(\varphi_1) - E'(\varphi_0)}{2} > \left| \frac{E'(\varphi_1) + E'(\varphi_0)}{2} \right| \qquad \text{(B8)}$$

Assuming:

$$\int_0^1 w_\nu \; d\nu = 1$$

we can use Equation 7 to find that:

$$E'(\varphi_0) = -1 + 2 \int_0^{\varphi_0} d\nu \; w_\nu - \varphi_0 + \bar{\nu} \qquad \text{(B9)}$$

$$E'(\varphi_1) = 1 - 2 \int_{\varphi_1}^1 d\nu \; w_\nu - \varphi_1 + \bar{\nu} \qquad \text{(B10)}$$

Substituting Equations B9 and B10 in B8, we obtain:

$$\int_{\varphi_0}^{\varphi_1} d\nu(w_\nu - 0.5) > \left| \int_0^{\varphi_0} d\nu(w_\nu - 0.5) - \int_{\varphi_1}^1 d\nu(w_\nu - 0.5) - (0.5 - \bar{\nu}) \right| \qquad \text{(B11)}$$

We now define the saliency of the k-th salient/nonsalient/semi-salient interval, A_k, to be the signed area bounded between w_μ and 0.5, i.e.:

$$A_k = \int d\nu (w_\nu - 0.5)$$

with the integral taken over the corresponding interval. Note that $A_k > 0$ for salient intervals, $A_k < 0$ for nonsalient intervals, and $A_k = 0$ for semisalient intervals. The condition for the formation of an attractor over the salient interval k can now be written as:

$$A_k > \left| \sum_{j<k} A_j - \sum_{j>k} A_j - (0.5 - \bar{\nu}) \right| \tag{B12}$$

Condition for Formation of a Line Attractor Over a Semisalient Interval

Let $\varphi_0 < \varphi_1$ denote the edges of the semisalient interval. For $\varphi_0 < \mu < \varphi_1$, the first derivative of the energy function is:

$$E'(\mu) = \int_0^{\varphi_0} d\nu (w_\nu - 0.5) - \int_{\varphi_1}^1 d\nu (w_\nu - 0.5) - (0.5 - \bar{\nu}) \tag{B13}$$

which gives some constant value that does not depend on μ. For a line attractor to be formed in the semisalient interval, this value has to be zero. Using the definition of A_k, we can write this condition as:

$$0 = \sum_{j<k} A_j - \sum_{j>k} A_j - (0.5 - \bar{\nu}) \tag{B14}$$

C. Analysis of Learning with the Gradual Protocol

In this section we analyze the dynamics of the attractor landscape at a single training session of the gradual protocol. We will assume that the weights acquired prior to the training are negligible, and we will consider the limit of a large number of morph patterns. Let w_ν^ψ denote the weight of pattern ν after the training with pattern ξ^ψ. Since every weight is updated once (after the presentation of the corresponding pattern), and using the assumption of negligible weights prior to training, we can write w_ν^ψ as:

$$w_\nu^\psi = \begin{cases} W_\nu & \nu \leq \psi \\ 0 & \nu > \psi \end{cases} \tag{C1}$$

where $W_\nu = w_\nu^1$ is the weight function at the end of the training. The value of W_ν is determined by the novelty-based learning rule. We will consider the version of the learning rule that involves the Hamming distance between the input and the attractor to which it is drawn (the analysis for the learning rule involving the distance between the input and the state after one update step is very similar). With appropriate normalization of the Hamming distance, W_ψ can be written as $W_\psi = \eta|\psi - \mu(\psi)|$, where $\mu(\psi)$ is the attractor to which ξ^ψ is drawn after it is presented. Because the weights that determine $\mu(\psi)$ vanish at values larger that ψ, any salient interval must be at the interval at $[0,\psi]$; therefore, $\mu(\psi) \leq \psi$, and we can write W_ψ as:

$$W_\psi = \eta(\psi - \mu(\psi)) \tag{C2}$$

$\mu(\psi)$ must satisfy the fixed point equation (Equation 5), i.e.:

$$\int_0^{\mu(\psi)} d\nu \; w_\nu^\psi (1 - |\nu - \mu(\psi)|) = \int_{\mu(\psi)}^1 d\nu \; w_\nu^\psi (1 - |\nu - \mu(\psi)|) \tag{C3}$$

Substituting Equation C1 in Equation C3, and with some rearrangement, we find that:

$$2 \int_0^{\mu(\psi)} d\nu \; W_\nu - \int_0^\psi d\nu \; W_\nu - \mu(\psi) \int_0^\psi d\nu \; W_\nu + \int_0^\psi d\nu \; W_\nu \nu = 0 \tag{C4}$$

The problem is now reduced to finding two functions, W_ψ and $\mu(\psi)$, that satisfy Equations C2 and C4 for any ψ. We achieve this by combining analytical and numerical methods.

First, note that by substituting W_ψ in Equation C4 with the right-hand side of Equation C2, it is clear that the solution for $\mu(\psi)$ does not depend on η. Thus, we will assume $\eta = 1$. Observe now that since $0 \leq \mu(\psi) \leq \psi$, it holds that $\mu(0) = 0$, and consequently, $W_0 = 0$. Next, we write power series expansions for W_ψ and $\mu(\psi)$:

$$W_\nu = \sum_{i=1}^\infty a_i \nu^i$$

$$\mu(\nu) = \sum_{i=1}^\infty a_i' \nu^i$$

where:

$$a_i' = \begin{cases} 1 - a_i & i = 1 \\ -a_i & i > 1 \end{cases} \tag{C5}$$

Substituting these expansions in Equation C4, we find that:

$$2 \sum_{i=1}^\infty \frac{a_i}{i+1} \left(\sum_{j=1}^\infty a_j' \psi^j \right)^{i+1} - \sum_{i=1}^\infty \frac{a_i}{i+1} \psi^{i+1} - \left(\sum_{i=1}^\infty \frac{a_i}{i+1} \psi^{i+1} \right) \times \left(\sum_{j=1}^\infty a_j' \psi^j \right) + \sum_{i=1}^\infty \frac{a_i}{i+2} \psi^{i+2} = 0 \tag{C6}$$

Collecting terms by the power of ψ gives:

$$\left(a_1 a_1'^2 - \frac{a_1}{2} \right) \psi^2 + \sum_{i=3}^\infty \left(\frac{2a_{i-1}a_1'^i}{i} + 2a_1 a_1' a_{i-1}' + \sum_{j=1}^{i-2} \frac{2a_j}{j+1} \Omega_i^\psi \left(\left(\sum_{k=1}^{i-2} a_k' \psi'^k \right)^{j+1} \right) - \frac{a_{i-1}}{i} - \sum_{j=1}^{i-2} \frac{a_{i-1-j}' a_j}{j+1} + \frac{a_{i-2}}{i} \right) \psi^i = 0 \tag{C7}$$

where $\Omega_i^\psi (P(\psi))$ returns the coefficient of ψ^i in a polynomial $P(\psi)$. For the function of ψ on the left-hand side of Equation C7 to vanish for all ψ, the coefficients of ψ^i must vanish for all i. Equating the coefficient of ψ^2 to zero, we get three solutions for the pair a_1 and a_1', out of which only one solution is positive for both a_1 and a_1' (a property required to guarantee that W_ψ and $\mu(\psi)$ are positive). The solution of a_1 is:

$$a_1 = 1 - \frac{\sqrt{2}}{2} \tag{C8}$$

Equating the coefficients of powers higher than 2 to zero, we find that:

$$a_i = \frac{a_{i-1} + (i+1)\sum_{j=1}^{i-1} \frac{a_j}{j+1} \left(2\Omega_{i+1}^\psi \left(\left(\sum_{k=1}^{i-1} a_k' \psi^k \right)^{j+1} \right) - a_{i-j}' \right)}{1 + 2(i+1)a_1 a_1' - 2a_1'^{i+1}} \tag{C9}$$

for $i \geq 2$. This equation for a_i depends only on $a_1, \ldots, a_{i-1}$ and $a_1', \ldots, a_{i-1}'$. Thus, using Equations C5, C8, and C9, it is straightforward to calculate recursively all a_i and a_i', and consequently obtain a power series approximation of W_ν and $\mu(\psi)$ up to an arbitrary power.

Performing these calculations, we find that the power series approximation converges quickly (with the first linear term being within 3% deviation from the exact solution; see Figure 6A). For the purpose of the analysis presented in the subsection "Protocol-Dependent Dynamics of Merging and Splitting of Attractors," we use this power series approximation to confirm that W_ν remains monotonic throughout the interval [0,1], which in turn implies that there is there is only one attractor throughout the training process.

Acknowledgments

References

Amit, D.J. (1995). The Hebbian paradigm reintegrated: local reverberations as internal representations. Behav. Brain Sci. *18*, 617–657.

Amit, D.J., and Brunel, N. (1997). Model of global spontaneous activity and local structured delay activity during delay periods in the cerebral cortex. Cereb. Cortex *7*, 237–252.

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

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

Bear, M.F., and Malenka, R.C. (1994). Synaptic plasticity: LTP and LTD. Curr. Opin. Neurobiol. *4*, 389–399.

Ben-Yishai, R., Bar-Or, R.L., and Sompolinsky, H. (1995). Theory of orientation tuning in visual cortex. Proc. Natl. Acad. Sci. USA *92*, 3844–3848.

Bliss, T.V.P., and Lomo, T.J. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. J. Physiol. *232*, 331–356.

Blumenfeld, B., Bibitchkov, D., and Tsodyks, M. (2006). Neural network model of the primary visual cortex: from functional architecture to lateral connectivity and back. J. Comput. Neurosci. *20*, 219–241.

Brunel, N. (2005). Network models of memory. In Methods and Models in Neurophysics, Volume Session LXXX: Lecture Notes of the Les Houches Summer School 2003, C.C. Chow, B. Gutkin, D. Hansel, C. Meunier, and J. Dalibard, eds. (Amsterdam: Elsevier), pp. 407–476.

Carpenter, G.A., and Grossberg, S. (2003). Adaptive resonance theory. In The Handbook of Brain Theory and Neural Networks Second Edition, M.A. Arbib, ed. (Cambridge, MA: MIT Press), pp. 87–90.

Chafee, M.V., and Goldman-Rakic, P.S. (1998). Matching patterns of activity in primate prefrontal area 8a parietal area 7ip neurons during a spatial working memory task. J. Neurophysiol. *79*, 2919–2940.

Deneve, S., Latham, P., and Pouget, A. (1999). Reading population codes: a neural implementation of ideal observers. Nat. Neurosci. *2*, 740–745.

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

Ermentrout, B. (1998). Neural networks as spatio-temporal pattern-forming systems. Rep. Prog. Phys. *61*, 353–430.

Fransen, E., Tahvildari, B., Egorov, A.V., Hasselmo, M.E., and Alonso, A.A. (2006). Mechanism of graded persistent cellular activity of entorhinal cortex layer V neurons. Neuron *49*, 735–746.

Friedman, D., Cycowicz, Y.M., and Gaeta, H. (2001). The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. Neurosci. Biobehav. Rev. *25*, 355–373.

Fuster, J.M., and Alexander, G. (1971). Neuron activity related to short-term memory. Neuron *14*, 477–485.

Gu, Q. (2002). Neuromodulatory transmitter systems in the cortex and their role in cortical plasticity. Neuroscience *111*, 815–835.

Hasselmo, M.E., and Schnell, E. (1994). Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: computational modeling and brain slice physiology. J. Neurosci. *14*, 3898–3914.

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

Hebb, D.O. (1949). The Organization of Behavior: A Neuropsychological Theory (New York: Wiley).

Hinton, G.E. (2002). Training products of experts by minimizing contrastive divergence. Neural Comput. *14*, 1771–1800.

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

Kirchhoff, B.A., Wagner, A.D., Maril, A., and Stern, C.E. (2000). Prefrontal-temporal circuitry for episodic encoding and subsequent memory. J. Neurosci. *20*, 6173–6180.

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

Li, S., Cullen, W.K., Anwyl, R., and Rowan, M.J. (2003). Dopamine-dependent facilitation of LTP induction in hippocampal CA1 by exposure to spatial novelty. Nature Neuroscience *6*, 526–531.

Lisman, J.E., and Grace, A.A. (2005). The hippocampal-VTA loop: controlling the entry of information into long-term memory. Neuron *46*, 703–713.

McNaughton, B.L., and Morris, R.G.M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. Trends Neurosci. *10*, 408–415.

Meltzer, J.A., and Constable, R.T. (2005). Activation of human hippocampal formation reflects success in both encoding and cued recall of paired associates. Neuroimage *24*, 384–397.

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

Miyashita, Y., and Chang, H.S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. Nature *331*, 68–70.

Movellan, J. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In Proceedings of the 1990 Connectionist Models Summer School, D. Touretzky, J. Elman, T. Sejnowski, and G. Hinton, eds. (San Mateo, CA: Morgan Kaufmann), pp. 10–17.

Nyberg, L. (2005). Any novelty in hippocampal formation and memory? Curr. Opin. Neurol. *18*, 424–428.

Otani, S., Daniel, H., Roisin, M.P., and Crepel, F. (2003). Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. Cereb. Cortex *13*, 1251–1256.

Preminger, S., Sagi, D., and Tsodyks, M. (2005). Morphing visual memories through gradual associations. Perception Suppl. *34*, 14.

Ranganath, C., and Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. Nat. Rev. Neurosci. *4*, 193–202.

Samsonovich, A., and McNaughton, B.L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. J. Neurosci. *17*, 5900–5920.

Schultz, W. (2002). Getting formal with dopamine and reward. Neuron *36*, 241–263.

Seung, H.S. (1996). How the brain keeps the eyes still. Proc. Natl. Acad. Sci. USA *93*, 13339–13344.

Treves, A. (1990). Graded-response neurons and information encodings in autoassociative memories. Phys. Rev. A. *42*, 2418–2430.

Treves, A., and Rolls, E.T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. Hippocampus *2*, 189–199.

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

Tsodyks, M., and Sejnowski, T. (1995). Associative memory and hippocampal place cells. Int. J. Neural. Syst. *6* (Suppl. 1995), 81–86.

Tsodyks, M.V., and Feigel'man, M.V. (1988). The enhanced storage capacity in neural networks with low level of activity. Europhys. Lett. *6*, 101–105.

Tulving, E., Markowitsch, H.J., Craik, F.E., Habib, R., and Houle, S. (1996). Novelty and familiarity activations in PET studies of memory encoding and retrieval. Cereb. Cortex *6*, 71–79.

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