LETTER **Communicated by Rafal Bogacz**

Dynamics and Robustness of Familiarity Memory

J. M. Cortes
jcortes1@inf.ed.ac.uk
A. Greve
grevea@cardiff.ac.uk
A. B. Barrett
abarrett@inf.ed.ac.uk
M. C. W. van Rossum
mvanross@inf.ed.ac.uk
Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh Informatics Forum, Edinburgh, EH8 9AB, U.K.

When presented with an item or a face, one might have a sense of recognition without the ability to recall when or where the stimulus has been encountered before. This sense of recognition is called *familiarity memory*. Following previous computational studies of familiarity memory, we investigate the dynamical properties of familiarity discrimination and contrast two different familiarity discriminators: one based on the energy of the neural network and the other based on the time derivative of the energy. We show how the familiarity signal decays rapidly after stimulus presentation. For both discriminators, we calculate the capacity using mean field analysis. Compared to recall capacity (the classical associative memory in Hopfield nets), both the energy and the slope discriminators have bigger capacity, yet the energy-based discriminator has a higher capacity than one based on its time derivative. Finally, both discriminators are found to have a different noise dependence.

1 Introduction

It is believed that recognition memory is supported by at least two different types of retrieval processes: recollection and familiarity (for a review, see Yonelinas, 2002). While recollection requires detailed information about an experienced event, familiarity just distinguishes whether the stimulus was previously encountered. A well-known example is the encounter with a colleague during a conference: one might recognize the person but fail to remember the time and place of an earlier meeting.

Familiarity memory is thought to have a very large capacity. Standing (1973) tested the capacity in humans by presenting participants with a large number (10,000) of images. After just one presentation (i.e., one-shot learning), participants were able to successfully recognize most of the previously

 © 2009 Massachusetts Institute of Technology

seen pictures. It is this type of familiarity that we model, in contrast to neocortical models with slowly developing familiarity (Norman & O'Reilly, 2003).

It appears that the medial temporal lobe, in addition to the prefrontal cortex, plays a critical role in familiarity memory. One patient with an intact prefrontal cortex but impaired medial temporal lobe revealed severe deficits in familiarity processing (Bowles et al., 2007). (For recent reviews on the role of the medial temporal lobe in familiarity discrimination, including neuroimaging results, see Eichenbaum, Yonelinas, & Ranganath, 2007; Mayes, Montaldi, & Migo, 2007). Within the medial temporal lobe, it seems that different brain areas are engaged during recollection and familiarity processing (Brown & Aggleton, 2001). Single-item familiarity is believed to be processed in the perirhinal cortex, whereas recollection is believed to involve the hippocampus. Indeed, electrophysiological studies using single-cell recordings in monkeys and rats (Brown, Wilson, & Riches, 1987; Brown & Xiang, 1998) report that about 30% of neurons in the perirhinal cortex show increased activity on presentation of a novel as opposed to an old stimulus. These neurons have been interpreted as novelty detectors and could form the basis for familiarity memory.

The association between memory processes and brain area, however, is still somewhat unclear and seems to depend on the nature of the stimulus (Aggleton & Brown, 2005; Rugg & Yonelinas, 2003). For instance, Xiang and Brown (2004) reported greater neuronal response in the prefrontal cortex for old as opposed to novel stimuli, suggesting that familiarity processing might be supported by prefrontal regions, while novelty detection is associated with the medial temporal lobe (in particular, the perirhinal cortex).

Familiarity and recollection memory have distinct temporal characteristics. In neuroimaging studies using event-related potentials (ERPs), familiarity is linked to a frontal ERP modulation that occurs around 300 to 500 ms after stimulus presentation, while recollection evokes a parietal ERP modulation 500 to 800 ms after stimulus presentation (Rugg et al., 1998; Rugg & Yonelinas, 2003; Greve, Sterratt, Donaldson, Willshaw, & van Rossum, 2009). Hence, the speed of processing of familiarity discrimination is faster than recollection. Behavioral experiments provide further evidence for the difference in timing. If only limited time is allowed for a recognition decision, subjects rely primarily on familiarity rather than recollection (Dosher, 1984).

In computational neuroscience, modeling of recollection via attractor neural networks has a long history using auto-associator Hopfield networks (Hopfield, 1982; Amit, 1989). It is only more recently that familiarity discrimination has been studied (Bogacz & Brown, 2003; Metter, Myers, & Gluck, 2005; Yakovlev, Amit, Romani, & Hochstein, 2008; Greve et al., 2009). It has been found that the capacity for familiarity discrimination in associative memory networks is much greater than that for recollection. Under a wide range of conditions, familiarity capacity is proportional to

the number of synapses within the network (Bogacz & Brown, 2003; Greve et al., 2009), whereas the capacity for recollection is merely proportional to the square root of the number of synapses (i.e., the number of neurons in a fully connected network; Amit, 1989). Intuitively this difference in capacity is easily understood. Familiarity memory requires just a single bit per pattern (familiar versus nonfamiliar), whereas recollection requires retrieval of the whole pattern (pattern completion).

This letter has the following related objectives: (1) to study the dynamics of familiarity discrimination, which potentially could correlate the model to the above findings concerning the timing of familiarity; (2) to explore how well time derivative of the energy, or slope, discriminates familiarity (this familiarity measure was originally suggested by Hopfield, 1982, but has not been investigated since); (3) to calculate the capacity using a mean field analysis as has been done for recollection capacity in Hopfield nets; and (4) to analyze how neural noise affects familiarity discrimination.

This letter is organized as follows. After introducing the network, we compare two different familiarity discriminators: one based on the energy, previously introduced by Bogacz and Brown (2003), and one based on the slope of the energy. We find that the signal from both familiarity discriminators decays quickly after exposure to the stimulus. We then investigate the robustness to noise of familiarity detection by studying the effects of random fluctuations in the network activity. Finally, using a mean field analysis, we compute the storage capacity for both discriminators and find that the energy-based discriminator always outperforms the one based on its time derivative. Only in the limit of high noise do they perform equally well.

2 Network Setup

We consider a network of N binary neurons, each with activity $s_i(t) = \pm 1$, the two states corresponding respectively to firing and not firing. The complete network activity is characterized by the vector $\mathbf{s}(t)$. Any two neurons are connected by synaptic weights w_{ij}. As is standard in artificial network models (Amari, 1972; Hopfield, 1982), the network has a learning phase in which it encodes M stimuli $\mathbf{x}^\rho \equiv \{x_i^\rho\}_{i=1}^N$, $(\rho = 1, \ldots, M)$, in its weights using a Hebbian learning rule:

$$w_{ij} = \frac{1}{N} \sum_{\rho=1}^{M} x_i^\rho x_j^\rho. \tag{2.1}$$

It can be shown that of all local additive learning rules, rule 2.1 is optimal, as it provides the highest capacity in the limit of large N, M (Greve et al., 2009). During the subsequent test phase, the network's performance is evaluated. At $t = 0$, either an old (learned) or new (novel) probe stimulus

$\hat{\rho}$ is loaded into the network, $\mathbf{s}(t=0) = \mathbf{x}^{\hat{\rho}}$. Next, the stimulus is removed, and the network evolves freely.

The Hopfield network dynamics assumes that each neuron is updated precisely once, probabilistically and asynchronously, in each unit of time. (The biological duration that a time unit in the model corresponds to is hard to extract by comparing the model to, say, ERP data, given the additional delays present in biology, but it probably is about 10–100 ms.) As standard in artificial neural networks and in analogy with magnetic systems in physics, random fluctuations are included through a temperature parameter T. These so-called Glauber dynamics have been extensively studied in many different stochastic systems (Marro & Dickman, 1999). After the update, the probability distribution of the neuron's activity is

$$P\{s_i(t+1) = \pm 1\} = \frac{1}{1 + \exp[\mp 2\beta h_i(t)]}, \tag{2.2}$$

where $\beta \equiv 1/T$ is the inverse temperature parameter, and $h_i(t) \equiv \sum_{j=1}^{N} w_{ij} s_j(t)$ is the total synaptic current received by neuron i. Accordingly, for low temperature, the noise is small, and there is a strong correlation between the input current h_i and the output s_i, while for high temperature, the output of a node is dominated by noise, and as $T \to \infty$, the output is independent of its input.

The energy in the network at time t is defined as

$$E(t) \equiv -\sum_{ij} w_{ij} s_i(t) s_j(t). \tag{2.3}$$

In the absence of noise (zero temperature), the energy can only decrease or stay the same, so ultimately, the activity reaches the attractor state that corresponds to a memory. The energy can be thought of as a measure of the correlation between input to a neuron and its output activity, with greater correlation corresponding to lower energy. This can be seen by rewriting the energy in terms of the inputs h_i and the outputs s_i, yielding $E(t) = -\sum_i h_i(t) s_i(t)$.

Equation 2.3 also suggests a network that reads out the energy. One could construct an additional set of neurons that each calculate the product of h_i and s_i, and their activities are then summed in an output neuron to yield the energy. Although this is not a very elegant solution, as it requires a multiplication operation and a duplication of the synaptic weights, it does show that the network energy is not a purely theoretical quantity. (For other network implementations that read out the network energy, see, e.g., Bogacz, Brown, & Giraud-Carrier, 2001; Greve et al., 2009.) The time derivative of the energy can be easily calculated in neural circuits

once the energy has been extracted, for instance, using short-term synaptic depression (Puccini, Sanchez-Vives, & Compte, 2007).

3 Two Familiarity Discriminators

The energy $E(t)$ at time $t = 0$, can be used to discriminate between old and new stimuli (Bogacz & Brown, 2003). As shown below, the energy is initially of order $-(N + M)$ for old stimuli and of order $-M$ for new stimuli. Because the energies differ by order N, while the standard deviation is $\sqrt{2M}$, they are macroscopically different. We call the discriminator that calculates the difference in energy between old and new patterns *FamE*.

The time derivative, or slope, of the energy $S(t) = \frac{dE(t)}{dt}$ can also be used as a familiarity discriminator. It indicates how quickly the network's energy changes immediately after a stimulus is presented. Interestingly, this familiarity measure was originally proposed in Hopfield's seminal paper (Hopfield, 1982), but to the best of our knowledge, it has never received further exploration. We call the discriminator that calculates the difference in the slopes for old and new patterns *FamS*.

We express the energy and its time derivative as functions of the M-dimensional vector $\mathbf{m}(t) \equiv \{m^\rho(t)\}_{\rho=1}^{M}$. Its components are the overlaps between the current network activity and each of the stored patterns and are defined by

$$m^\rho(t) \equiv \frac{1}{N} \sum_{i=1}^{N} x_i^\rho s_i(t). \tag{3.1}$$

Assuming the Hebbian learning rule 2.1, the energy equation 2.3 in terms of the overlaps is

$$E(t) = -N \sum_{\rho=1}^{M} [m^\rho(t)]^2 , \tag{3.2}$$

while the time derivative of the energy is given by

$$S(t) = -2N \sum_{\rho=1}^{M} m^\rho(t) \frac{dm^\rho(t)}{dt}, \tag{3.3}$$

and is thus proportional to the time derivative $dm^\rho(t)/dt$ of the overlaps.

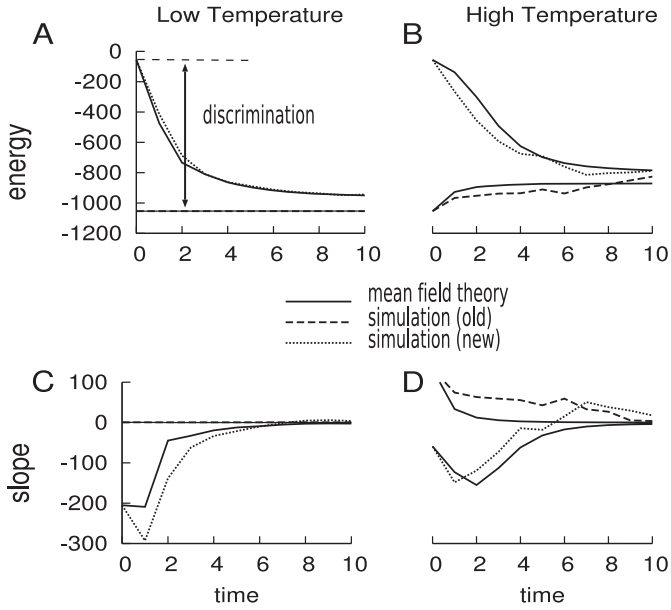

Figure 1: Temporal profile of familiarity discrimination. Simulation of a network with $N = 1000$ neurons storing $M = 50$ uncorrelated patterns for different values of the temperature, $T = 0.20$ on the left (A, C) and $T = 0.60$ on the right (B, D). Both the energy (A–B) and the slope (C–D) can discriminate between new and old stimuli during a short period after stimulus presentation. In graphs C–D, the slope rapidly tends to zero, indicating that the activity has converged to one of the stored stimuli. This is due to the well-known pattern completion dynamics that occurs in attractor neural networks. Solid lines correspond to the mean field theory. Dashed and dotted lines correspond to simulations for old and new patterns, respectively (note that the theory and simulation for old patterns overlap very closely). One unit of time is defined as the time taken to update all neurons in the network once.

4 Dynamics of the Familiarity Discriminators

We compared the two discriminators, FamE and FamS, in simulations of networks with Glauber dynamics, equation 2.2. The energy associated with old stimuli is initially much lower than for new stimuli (see Figures 1A and 1B). However, after a short transient of some five time units, the two signals become similar, that is, familiarity discrimination based on energy deteriorates rapidly after stimulus presentation as the energy associated with new and old stimuli becomes of the same order. The underlying reason is that the activity in the Hopfield network will always reach an attractor state regardless of the initial activity pattern. As the energy of the different attractors is

similar, the signal-to-noise ratio (below) is low and the discrimination poor. Small differences in the energy can remain for low levels of noise ($T = 0.20$ in Figure 1A), but they tend to reduce for high noise ($T = 0.6$ in Figure 1B). In the next section, we specifically study how the discrimination is affected by the noise parameter T.

Like the energy, its derivative also shows a transient signal when the network is presented with a new rather than an old stimulus (see Figures 1C and 1D). For low temperature, the slope for old stimuli is practically zero. This can be easily understood. An old stimulus corresponds to one of the local minima (attractors) of the energy landscape. At low temperature, the system does not receive any external perturbation, and so the energy does not change; its time derivative is zero. The derivative associated with old and new stimuli shows significant differences immediately after stimulus presentation, but this diminishes shortly after. Whatever the stimulus, the slope tends to zero as time progresses because the network evolves toward a fixed point and becomes stationary.

To mathematically address the network dynamics, we assume the mean field approximation: $s_i \approx \langle s_i \rangle$. Under this approximation, one obtains from equation 2.2 the dynamical equations for the overlaps

$$\frac{dm^{\rho}(t)}{dt} = -m^{\rho}(t) + \frac{1}{N}\sum_{i=1}^{N} x_i^{\rho} \tanh\left[\beta \sum_{\nu=1}^{M} x_i^{\nu} m^{\nu}(t)\right]. \tag{4.1}$$

The mean field formulation provides an accurate description of the dynamics of the system provided the temperature is not too high (see below). Indeed the theory matches the simulation well (see Figure 1, solid lines).

In summary, both the FamE and FamS discriminators distinguish old from new stimuli, but after a short transient of the order of five time units, discrimination ability of both discriminators disappears.

5 Robustness of the Familiarity Discriminators to Noise

Next we study how the temperature parameter, which quantifies random fluctuations in neural activity, affects the performance of the familiarity discriminators. We study the effect of temperature at two different time points, $t = 0$ and $t = 1$. Time is defined such that in one unit, all neurons are asynchronously updated once. The choice of $t = 1$ is not special; it is a convenient value somewhere between the initial and steady state.

Immediately after presentation of the stimulus, the energy is independent of temperature (see Figure 2A). The reason is that by definition, the energy at $t = 0$ will be calculated before any activity updates have occurred. In contrast, the slope has a nonlinear relationship with temperature (see Figure 2C) and interestingly performs best as a familiarity discriminator at a high rather than a low temperature. The slope is proportional to the rate of

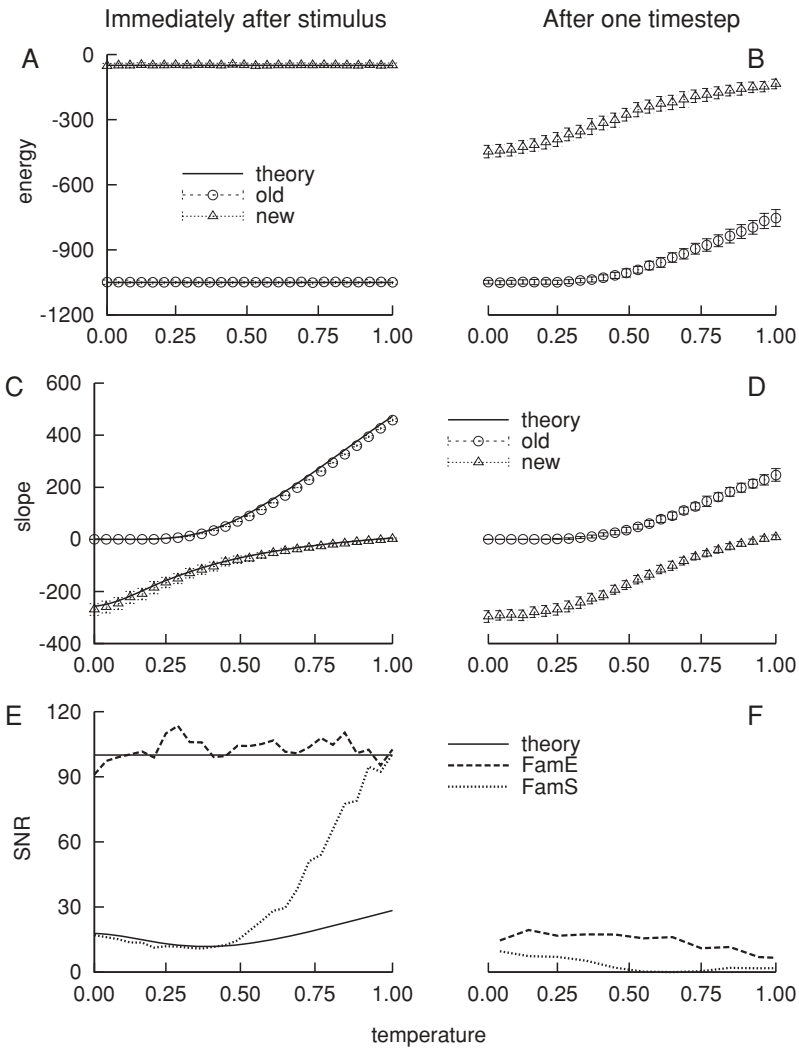

Figure 2: Robustness of familiarity discrimination to noise. Immediately after stimulus presentation, the energy is independent of temperature for both old (circles) and new (triangles) stimuli (A), whereas the slope is temperature dependent (C). After one time step, both energy (B) and slope (D) are temperature dependent. Circles and triangles (shown with standard deviation) represent responses to old and new stimuli respectively. The bottom row shows the SNR of the familiarity discriminators against temperature. Simulations averaging over 100 runs of a network with $N = 1000$ neurons and $M = 50$ stored patterns. Only for graph F is the number of runs 500. Black solid lines are the theoretical predictions (not available for $t = 1$; see the text).

change of the overlap between the network activity and the stimulus. At low temperatures, the slope associated with an old stimulus is approximately zero, as the overlap with the stimulus is almost invariant. Contrarily, at high temperature, the overlap with old stimuli changes very quickly. It decays approximately from 1 to 0 (the fully disordered state), and consequently the slope is high. As a result, FamS performs better at higher temperatures. Note that familiarity discrimination is still possible for $T > 1$, but recollection is not. For $T > 1$ the only stable solution is $\mathbf{m} = 0$, the so-called *paramagnetic* or *nonmemory* solution in associative networks (Amit, 1989). We do not study this regime because the initial condition $m = 1$ after stimulus presentation is inconsistent with the stationary paramagnetic solution $m \approx 0$.

In contrast to time $t = 0$, at time $t = 1$, both discriminators show a similar breakdown in performance, in particular at increased temperature (see Figures 2B, 2D, and 2F). Our measure for performance is defined through the signal-to-noise ratio, which is introduced in the next section. In conclusion, in particular at $t = 0$, the FamE and FamS discriminators work well, but the slope works best at high noise.

6 Storage Capacity

To examine the capacity of the two familiarity discriminators, we quantify the discriminability between their responses to new and old stimuli by using the signal-to-noise ratio (SNR). The SNR for FamE is defined as

$$\mathrm{SNR(FamE)} = \frac{|\langle E_{\mathrm{new}}\rangle - \langle E_{\mathrm{old}}\rangle|}{\sqrt{\frac{1}{2}\mathrm{Var}(E_{\mathrm{new}}) + \frac{1}{2}\mathrm{Var}(E_{\mathrm{old}})}} \tag{6.1}$$

and is analogous for the slope. The mean and variances are computed averaging over many different configurations of patterns. In general, the energy and slope distributions associated with both old and new stimuli are well described by gaussians. Numerically (using 100 trials), the fourth moment satisfies within 5% that $\langle x^4\rangle = \int P(x)x^4 dx = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$, where $\mu = \langle x\rangle$ denotes the mean and $\sigma^2 = \langle x^2\rangle - \langle x\rangle^2$ the variance. In particular at low temperatures, the slope distribution associated with old stimuli starts to deviate from a gaussian. In this case, the slope is often zero, and sometimes positive. As a result, the distribution more sharply peaked at zero and has a positive skew.

When the number of stimuli encoded in the weights increases, the SNR decreases. We define the storage capacity (or maximum number of stimuli encoded in the learning rule and successfully discriminated) as the point where the SNR drops below some constant, say unity. This gives the maximum number of stimuli M_{max} that can be encoded in the network. Next, we present analytical calculations for the capacity of both discriminators at time $t = 0$.

6.1 Storage Capacity of FamE, the Energy Discriminator. Let $\rho = \hat{\rho}$ label an old stimulus presented to the network. As is common in these calculations (Hertz, Krogh, & Palmer, 1991), we separate the sum appearing in equation 3.2 into signal ($\rho = \hat{\rho}$) and noise ($\rho \neq \hat{\rho}$) contributions. At $t = 0$ and for many neurons N, applying the overlap definition, equation 3.1, yields that the overlap $m^{\hat{\rho}}$ has mean 1 and variance 0. The overlaps with $\rho \neq \hat{\rho}$ have mean 0 and variance $1/N$. Thus, the noise term in equation 3.2 can be written using a χ^2 with $M-1$ degrees of freedom with mean $(M-1)$ and variance $2(M-1)$. With the patterns $m^{\hat{\rho}}$ and $m^{\rho\neq\hat{\rho}}$ uncorrelated, one finds for large M, $\langle E_{\text{old}}\rangle = -(N+M)$ and $\text{Var}(E_{\text{old}}) = 2M$. Analogously, but with no signal term, the energy for new stimuli satisfies $\langle E_{\text{new}}\rangle = -M$ and $\text{Var}(E_{\text{new}}) = 2M$. Directly from equation 6.1, we obtain $\text{SNR} = \sqrt{N^2/(2M)}$, in agreement with our simulations, Figure 2E. The storage capacity is found by solving $\text{SNR} = 1$ for M, which gives

$$M_{\max}[\text{FamE}] = \frac{N^2}{2}, \tag{6.2}$$

and thus the storage is of order N^2, as found in previous models using the energy discriminator (Bogacz & Brown, 2003; Greve et al., 2009).

6.2 Storage Capacity of FamS, the Slope Discriminator. Directly by substitution of equation 4.1 in equation 3.3, the slope can be written as

$$S = 2(\hat{E} - E) \tag{6.3}$$

with $\hat{E}$ defined as

$$\hat{E} = -N \sum_{\rho=1}^{M} m^{\rho} \frac{1}{N} \sum_{i=1}^{N} \xi_i^{\rho} \tanh\left[\beta \sum_{\nu=1}^{M} x_i^{\nu} m^{\nu}(t)\right]. \tag{6.4}$$

From equation 6.3, the expected value is $\langle S\rangle = 2\langle\hat{E}\rangle - 2\langle E\rangle$ and the variance $\text{Var}(S) = 4\text{Var}(\hat{E}) + 4\text{Var}(E) - 8\text{Cov}(\hat{E}, E)$. As a first approximation, we will assume that both $\text{Var}(\hat{E})$ and $\text{Cov}(\hat{E}, E)$ are equal to zero. In this case, the only contribution to the variance of S comes from the variance of E. The mean value of $\hat{E}$ is computed in appendix A. For old stimuli, we obtain

$$\begin{aligned} \langle S_{\text{old}}\rangle &= 2N(1 - I_1 - I_2) + 2M \\ \text{Var}(S_{\text{old}}) &= 8M, \end{aligned} \tag{6.5}$$

where I_1 and I_2 are defined below, while for new stimuli,

$$\begin{aligned}\langle S_{\text{new}}\rangle &= -2NI_3 + 2M \\ \text{Var}(S_{\text{new}}) &= 8M.\end{aligned} \tag{6.6}$$

The integrals I_1, I_2, and I_3 are given by

$$\begin{aligned}I_1(\alpha,\beta) &\equiv \int \frac{dz}{\sqrt{2\pi}} \exp(-z^2/2)\tanh\left(\beta+\beta\sqrt{\alpha}z\right), \\ I_2(\alpha,\beta) &\equiv \int \frac{dz}{\sqrt{2\pi}} \exp(-z^2/2)\tanh\left(\beta+\beta\sqrt{\alpha}z\right)\sqrt{\alpha}z, \\ I_3(\alpha,\beta) &\equiv \int \frac{dz}{\sqrt{2\pi}} \exp(-z^2/2)\tanh\left(\beta\sqrt{\alpha}z\right)\sqrt{\alpha}z,\end{aligned} \tag{6.7}$$

where $\beta \equiv 1/T$ is the inverse temperature and $\alpha \equiv M/N$ is defined as the network load. From equations 6.5 and 6.6, it follows that

$$\text{SNR(FamS)} = \sqrt{N^2/(2M)}[1 - I_1(\alpha,\beta) - I_2(\alpha,\beta) + I_3(\alpha,\beta)]. \tag{6.8}$$

This is plotted against temperature in Figure 2E. For low temperature, there is good agreement with the simulation results, but for high temperatures, theory and simulation diverge. The theoretical mean values fit well with the simulations, Figure 2C, but the theoretical predictions for the variances of S_{old} and S_{new} are incorrect. In appendix B, we describe how the mean field approximation is affected by high temperatures.

The simulations we have presented thus far have a low network load ($\alpha = 0.05$), but familiarity discrimination remains possible for much larger values of α. In Figure 3 we store up to $M = 4000$ patterns in a network of $N = 1000$ neurons. For FamE, theory and simulations are in full agreement. For FamS, we observe a strong overestimation of the theory (see equation 6.8, curve SNR_1 in Figure 3) compared to simulation. The theoretical results were derived assuming that $\text{Var}(\hat{E})$ and $\text{Cov}(\hat{E}, E)0$ are very small. For large α, the approximation $\text{Var}(\hat{E}) \approx 0$ becomes invalid. Including corrections from $\text{Var}(\hat{E})$, we obtain at zero temperature,

$$\text{SNR}_2(\text{FamS}) = \frac{\text{SNR}_1(\text{FamS})}{\sqrt{1 + \frac{1}{4}[\alpha+1]\, I_4(\alpha,\beta) + \frac{1}{2}\alpha I_5(\alpha,\beta) + \frac{1}{4} I_6(\alpha,\beta)}}, \tag{6.9}$$

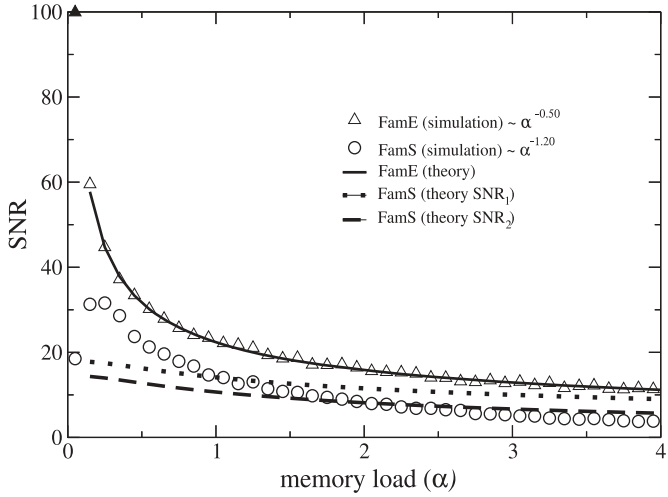

Figure 3: Effect of memory load on familiarity discrimination. The performance (SNR) decreases with increasing memory load ($\alpha = M/N$). Triangles (energy) and circles (slope) correspond to simulations of a network of $N = 1000$ neurons at zero temperature, 100 trials. The SNR of the energy-based discriminator scales as $\alpha^{-0.50}$ and the SNR of slope-based discriminator as $\alpha^{-1.2}$. The two curves, SNR_1 and SNR_2, correspond to two different approximations valid, respectively, for low and high α (see the text).

where SNR_1(FamS) is given by equation 6.8. The new integrals are

$$I_4(\alpha, \beta) \equiv \int \frac{dz}{\sqrt{2\pi}} \exp(-z^2/2) \tanh^2\left(\beta + \beta\sqrt{\alpha}z\right),$$

$$I_5(\alpha, \beta) \equiv \int \frac{dz}{\sqrt{2\pi}} \exp(-z^2/2) \tanh^2\left(\beta + \beta\sqrt{\alpha}z\right) \sqrt{\alpha}z,$$

$$I_6(\alpha, \beta) \equiv \int \frac{dz}{\sqrt{2\pi}} \exp(-z^2/2) \tanh^2\left(\beta\sqrt{\alpha}z\right). \tag{6.10}$$

In Figure 3, the curve SNR_1 corresponds to assuming $\mathrm{Var}(\hat{E}) = \mathrm{Cov}(\hat{E}, E) = 0$, that is, equation 6.8, and SNR_2 to the case of $\mathrm{Cov}(\hat{E}, E) = 0$, equation 6.9. The SNR_1 is valid for very low α (cf. 0.05 in Figure 2E), but it fails for intermediate and large values of α. The SNR_2 improves the prediction of the SNR for high loads of the network. Note that due to the existence of a critical point in the retrieval phase in Hopfield nets, mean field approximation of equation 4.1, and hence our results, fail around $\alpha \approx 0.14$ (Amit, Gutfreund, & Sompolinsky, 1987; Amit, 1989). In contrast to the theory, the SNR found in simulation peaks around this point.

To compute the capacity for FamS, we proceed similarly to FamE. We use the approximation of equation 6.9, which is valid for high α. The storage capacity for FamS is again obtained by solving SNR $= 1$, and this yields

$$M_{\max}[\text{FamS}] = \frac{N^2}{2} \frac{(1 - I_1(\alpha_{\max}, \beta) - I_2(\alpha_{\max}, \beta) + I_3(\alpha_{\max}, \beta))^2}{1 + \frac{1}{4}[\alpha + 1] I_4(\alpha, \beta) + \frac{1}{2}\alpha I_5(\alpha, \beta) + \frac{1}{4} I_6(\alpha, \beta)}. \tag{6.11}$$

This cannot readily be solved because the integrals depend on M through α. Interestingly, the capacity of FamS is also dependent on the temperature, while that of FamE is completely independent of temperature (recall Figures 2A and 2C).

In the two limits $T = 0$ and $T \to \infty$, we can solve the integrals in equation 6.11 to obtain the storage capacity. For $T = 0$, we use

$$\begin{aligned}
&\lim_{\beta\to\infty} \int \frac{dz}{\sqrt{2\pi}} \exp(-z^2/2) \tanh(\beta[az + b]) = \operatorname{erf}\left(\frac{b}{\sqrt{2}a}\right), \\
&\lim_{\beta\to\infty} \int \frac{dz}{\sqrt{2\pi}} \exp(-z^2/2) \tanh(\beta[az + b])\, z = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{b^2}{2a^2}\right) \\
&\lim_{\beta\to\infty} \tanh^2(\beta[az + b]) = 1,
\end{aligned} \tag{6.12}$$

giving $\lim_{\beta\to\infty} I_1(\alpha, \beta) = \operatorname{erf}(\frac{1}{\sqrt{2\alpha}})$, $\lim_{\beta\to\infty} I_2(\alpha, \beta) = \sqrt{\frac{2\alpha}{\pi}} \exp(-\frac{1}{2\alpha})$, $\lim_{\beta\to\infty} I_3(\alpha, \beta) = \sqrt{\frac{2\alpha}{\pi}}$, $\lim_{\beta\to\infty} I_4(\alpha, \beta) = 1$, $\lim_{\beta\to\infty} I_5(\alpha, \beta) = 0$ and $\lim_{\beta\to\infty} I_6(\alpha, \beta) = 1$, where $\operatorname{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x \exp(-u^2) du$ is the error function.

Thus, at $T = 0$, equation 6.11 becomes

$$M_{\max} = \frac{N^2}{2} \frac{\left(1 - \operatorname{erf}\left(\sqrt{\frac{N}{2M_{\max}}}\right) + \sqrt{\frac{2M_{\max}}{\pi N}}\left[1 - \exp\left(-\frac{N}{2M_{\max}}\right)\right]\right)^2}{1 + \frac{1}{4}\left[M_{\max}/N + 1\right] + \frac{1}{4}}. \tag{6.13}$$

Solving this self-consistent equation yields $M_{\max} \propto N^{3/2}$ as $N \to \infty$. This is smaller than storage achieved by the energy ($\propto N^2$), but it is still much higher than the recall capacity ($\propto N$).

Fitting the simulation results to a curve with form $SNR \propto \sqrt{N}\alpha^{-\gamma}$ yields $\gamma = 0.50$ and $\gamma = 1.2$, respectively. This corresponds to a capacity $M_{\max} \propto N^2$ for FamE and $M_{\max} \propto N^{1.42}$ for FamS, which is close to the analytical result. In Figure 4 we plot, as a function of N, the storage capacity ratio of FamS and FamE at zero temperature.

In the other limit that $T \to \infty$, random fluctuations in neural activity dominate the network dynamics. All the integrals in equations 6.7 and 6.10

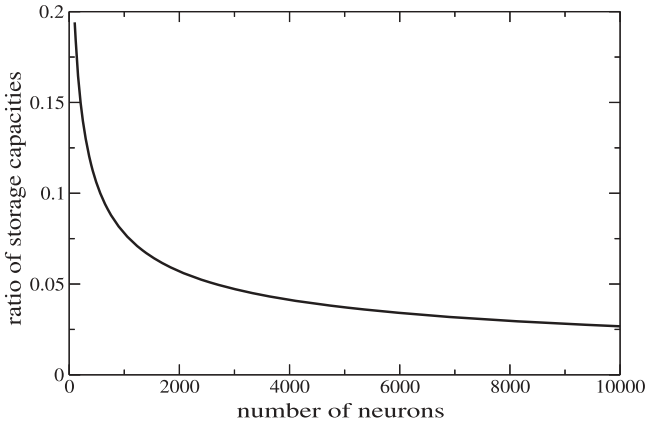

Figure 4: Ratio of storage capacities at zero temperature. The storage of the slope discriminator is obtained by numerical solution of equation 6.13 as a function of the number of neurons N. This is divided by the capacity of the energy-based discriminator storage, equation 6.2, to obtain the capacity ratio of the two discriminators.

are zero, and hence $M_{\max}[\text{FamS}] \approx M_{\max}[\text{FamE}] \approx N^2/2$. In this high-noise limit, the theoretical storage capacity is the same for both discriminators. For arbitrary temperatures, the capacity can be obtained by numerical evaluation of the integrals.

Unfortunately, the mean field analysis cannot be used for times other than $t = 0$. This regime would require more advanced techniques such as generating functional analysis (Coolen, 2001). Network simulations for $t = 1$ were shown in Figures 2B, 2D, and 2F.

7 Discussion

Familiarity describes a retrieval process that supports recognition memory. Numerous empirical studies have investigated familiarity processes in humans (Yonelinas, 2002) and mammals (Brown & Xiang, 1998). Recently neuronal networks modeling familiarity discrimination have been proposed (Bogacz & Brown, 2003; Yakovlev et al., 2008). This study extends these results in a number of directions: First, we analyzed an alternative familiarity discriminator, FamS. Second, we examined the dynamics of the familiarity signal and, finally, we show how familiarity memory can be analyzed in a mean field framework.

We have compared the energy discriminator used by Bogacz and Brown (2003) to a discriminator based on its time derivative, or slope. The latter discriminator was suggested by Hopfield (1982) in his seminal study,

but had not been explored before. Here we have shown that the slope works well as a familiarity discriminator and is a good indicator of whether the stimulus has been presented during learning or is novel. Thus, the same Hopfield network can be used for both recollection (stationary properties of the retrieval dynamics) and familiarity (transient dynamics after the stimulus presentation) (Greve, Donaldson, & van Rossum, in press).

For both discriminators, the signal decays quickly after stimulus presentation. This can be compared to the speed of recollection. Assuming that recollection memories correspond to reaching an attractor in the Hopfield model, recollection information becomes available only when activity has settled. By that time, the slope signal is zero, and the energy signal is also very weak (although not necessarily zero). Human familiarity is likely very complicated, and our model is an extreme simplification. As a result, it is hard to justify mapping our findings to experimental studies. Nevertheless, the experimentally observed timing difference between familiarity and recollection is consistent with our model.

The storage capacity for familiarity memory is always larger than the recall capacity of memories in Hopfield nets (proportional to the number of neurons N), consistent with the observed high capacity of familiarity memory (Standing, 1973). The capacity depends on the noise. In the low-noise limit, FamE has a storage proportional to N^2, and FamS has a capacity proportional to $N^{3/2}$. In the high-noise limit, the storage capacities of both FamE and FamS are approximately $N^2/2$. Interestingly, this means that the slope performance improves as one goes to the high-noise regime (see Figure 2E). This stands in stark contrast with how noise affects recollection in Hopfield nets, where noise decreases the recollection performance (Amit, 1989).

It is worth noting that we considered storage of only uncorrelated patterns. This means that the local memory attractors are deep and well separated (Amit, 1989). In simulation with correlated patterns, we found that the performance of both discriminators decreases similarly (not shown).

In this study, a single stimulus presentation of a stimulus during training is sufficient for a subsequent familiarity memory. In contrast to Norman and O'Reilly (2003) and Yakovlev et al. (2008), the model does need repeated presentation of stimuli to enable familiarity discrimination. Although the main purpose of this letter is not to explore how repeated stimuli presentation affects the familiarity performance, one could still use the synaptic matrix, equation 2.1. The effect of repeating a stimulus is simply to increase its energy in proportion to the number of repetitions. Thus, repeated stimuli will be more familiar than stimuli presented only once, and the strength of the memory can be used to distinguish whether a stimulus has been presented just once or many times, allowing for a flexible, high-capacity familiarity system.

Appendix A: Mean and Variance of $\hat{E}$

According to equation 6.3, $\hat{E}$ defined in equation 6.4 gives a relationship between slope and energy. Similar to Amit et al. (1987) expectations of $\hat{E}$ can be computed for large N, approximating the sum over the different sites i of the noise terms $\sum_{\rho\neq\hat{\rho}} x_i^{\rho} m^{\rho}$ appearing inside the tanh function with an integral over a gaussian measure with mean 0 and variance $\alpha = M/N$. Separating the signal ($\rho = \hat{\rho}$) from the noise ($\rho \neq \hat{\rho}$) in the case of presenting an old stimulus and with no signal for new stimuli, one obtains, after some algebra,

$$\begin{aligned}\langle \hat{E}_{\text{old}} \rangle &= -N\langle\langle \tanh(\beta + \beta\sqrt{\alpha}z)\rangle\rangle - N\langle\langle \sqrt{\alpha}z \tanh(\beta + \beta\sqrt{\alpha}z)\rangle\rangle \\ \langle \hat{E}_{\text{new}} &= -N\langle\langle \sqrt{\alpha}z \tanh(\beta\sqrt{\alpha}z)\rangle\rangle, \end{aligned} \tag{A.1}$$

where we have denoted $\langle\langle \mathrm{f}(z)\rangle\rangle \equiv \int \frac{dz}{\sqrt{2\pi}} \exp(-z^2/2)\mathrm{f}(z)$. To compute the variance, first one has to derive the second moment, squaring equation 6.4, and considering four different terms: $i = j$ & $\rho = \rho'$, $i = j$ & $\rho \neq \rho'$, $i \neq j$ & $\rho = \rho'$, and $i \neq j$ & $\rho \neq \rho'$. Separating signal and noise contributions, we obtain

$$\begin{aligned}\langle (\hat{E}_{\text{old}})^2 \rangle = &-\langle E_{\text{old}}\rangle \langle\langle \tanh^2\left(\beta + \beta\sqrt{\alpha}z\right)\rangle\rangle + N^2\langle\langle \tanh\left(\beta + \beta\sqrt{\alpha}z\right)\rangle\rangle^2 + \\ &+ 2N\langle\langle \sqrt{\alpha}z \tanh^2\left(\beta + \beta\sqrt{\alpha}z\right)\rangle\rangle + \\ &+ 2N^2\langle\langle \tanh\left(\beta + \beta\sqrt{\alpha}z\right)\rangle\rangle\langle\langle \sqrt{\alpha}z \tanh\left(\beta + \beta\sqrt{\alpha}z\right)\rangle\rangle + \\ &+ N^2\langle\langle \sqrt{\alpha}z \tanh\left(\beta + \beta\sqrt{\alpha}z\right)\rangle\rangle^2 \\ \langle (\hat{E}_{\text{new}})^2 \rangle = &-\langle E_{\text{new}}\rangle \langle\langle \tanh^2\left(\beta\sqrt{\alpha}z\right)\rangle\rangle + N^2\langle\langle \sqrt{\alpha}z \tanh\left(\beta\sqrt{\alpha}z\right)\rangle\rangle^2. \end{aligned} \tag{A.2}$$

Eventually, by the definition of *variance*, one gets

$$\begin{aligned}\mathrm{Var}(\hat{E}_{\text{old}}) = &-\langle E_{\text{old}}\rangle \langle\langle \tanh^2(\beta + \beta\sqrt{\alpha}z)\rangle\rangle + \\ &+ 2N\langle\langle \sqrt{\alpha}z \tanh^2(\beta + \beta\sqrt{\alpha}z)\rangle\rangle \\ \mathrm{Var}(\hat{E}_{\text{new}}) = &-\langle E_{\text{new}}\rangle \langle\langle \tanh^2(\beta\sqrt{\alpha}z)\rangle\rangle, \end{aligned} \tag{A.3}$$

which leads to SNR of FamS.

Appendix B: Temperature Dependence of Accuracy of Mean Field Approximation

To compute S from equation 3.3, we need an analytical expression for dm^{ρ}/dt, or equivalently, given the definition 3.1, we have to compute the

derivative ds_i/dt. Given $s_i(t)$, the Glauber dynamics give an uncertainty in $s_i(t+1)$ such that (Marro & Dickman, 1999)

$$\mathrm{Var}[s_i(t+1)|\{s_j(t)\}] = \mathrm{sech}^2(\beta h_i(t)), \tag{B.1}$$

which implies

$$\frac{ds_i}{dt} = \tanh(\beta h_i) - s_i + \mathcal{O}(\mathrm{sech}(\beta h_i)). \tag{B.2}$$

We use this result to find the error induced in our calculation of S_{new}. When a new pattern is presented, the m^ρ are all of order $N^{-1/2}$. This implies that the local fields, $h_i \equiv \sum_\rho x_i^\rho m^\rho$, are of order $\sqrt{\alpha}$. Hence, by equations 3.1 and B.2, the error in our calculation of dm^ρ/dt is given by

$$\mathrm{Error}\left(\frac{dm^\rho}{dt}\right) = \mathcal{O}\left(\frac{1}{\sqrt{N}}\mathrm{sech}\left(\beta\sqrt{\alpha}\right)\right), \tag{B.3}$$

for each ρ. Thus, by equation 3.3, we conclude that

$$\mathrm{Error}\,(S_{\mathrm{new}}) = \mathcal{O}\left(\sqrt{M}\,\mathrm{sech}\left(\frac{1}{T}\sqrt{\frac{M}{N}}\right)\right). \tag{B.4}$$

Since $\mathrm{sech}(x)$ decays exponentially with large x but is of order 1 for small x, the error in our calculation of S_{new}, coming from the mean field approximation, is going to be negligible only in the limit in which $(1/T)\sqrt{M/N}$ is large (i.e., low temperatures). The error in our calculation of S_{old} is similar. This analysis explains the growing discrepancy between theory and simulation as the temperature is increased (see Figure 2E).

Acknowledgments

We acknowledge Rafal Bogacz (University of Bristol) and David Donaldson (University of Stirling) for helpful discussions, the anonymous reviewers, and financial support from EPSRC (project EP/CO 10841/1), HFSP (project RGP0041/2006) and the Doctoral Training Center in Neuroinformatics at the University of Edinburgh.

References

Aggleton, J. P., & Brown, M. W. (2005). Contrasting hippocampal and perirhinal cortex function using immediate early gene imaging. *Q. J. Exp. Psychol. B*, *58*, 218–233.

Amari, S. I. (1972). Learning pattern sequences by self-organizing nets of threshold elements. *IEEE Trans. Comput. C*, *21*, 1197–1206.

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. Cambridge: Cambridge University Press.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1987). Statistical mechanics of neural networks near saturation. *An. Phys.*, *173*, 30–67.

Bogacz, R., & Brown, M. W. (2003). Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus*, *13*, 494–524.

Bogacz, R., Brown, M. W., & Giraud-Carrier, C. (2001). Model of familiarity discrimination in the perirhinal cortex. *J. Comput. Neurosci.*, *10*, 5–23.

Bowles, B., Crupi, C., Mirsattari, S. M., Pigott, S. E., Parrent, A. G., Pruessner, J. C., et al. (2007). Impaired familiarity with preserved recollection after anterior temporal-lobe resection that spares the hippocampus. *Proc. Natl. Acad. Sci. USA*, *104*, 16382–16387.

Brown, M. W., & Aggleton, J. P. (2001). Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nat. Rev. Neurosci.*, *2*, 51–61.

Brown, M. W., Wilson, F. A. W., & Riches, I. P. (1987). Neuronal evidence that inferomedial temporal cortex is more important than hippocampus in certain processes underlying recognition memory. *Brain Res.*, 409, 158–162.

Brown, M. W., & Xiang, J. Z. (1998). Recognition memory: Neuronal substrates of the judgment of prior occurrence. *Prog. Neurobiol.*, *55*, 149–189.

Coolen, A. C. C. (2001). Statistical mechanics of recurrent neural networks II: Dynamics. In F. Moss & S. Gielen (Eds.), *Handbook of biological physics* (pp. 597–622). Amsterdam: Elsevier Science.

Dosher, B. A. (1984). Discriminating preexperimental (semantic) from learned (episodic) associations: A speed-accuracy study. *Cogn. Psychol.*, *16*, 519–555.

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annu. Rev. Neurosci.*, 30, 123–152.

Greve, A., Donaldson, D. I., & van Rossum, M. C. W. (in press). A single-trace dual-process model of episodic memory: A novel computational account of familiarity and recollection. *Hippocampus*.

Greve, A., Sterratt, D. C., Donaldson, D. I., Willshaw, D. J., & van Rossum, M. C. W. (2009). Optimal learning rules for familiarity detection I. *Biol. Cyber.*, *100*, 11–19.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Reading, MA: Addison-Wesley Longman.

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

Marro, J., & Dickman, R. (1999). *Nonequilibrium phase transitions in lattice models*. Cambridge: Cambridge University Press.

Mayes, A., Montaldi, D., & Migo, E. (2007). Associative memory and the medial temporal lobes. *Trends. Cogn. Sci.*, *11*, 126–135.

Metter, M., Myers, C. E., & Gluck, M. A. (2005). Integrating incremental learning and episodic memory models of the hippocampal region. *Psychol. Rev.*, *112*, 560–585.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychol. Rev.*, *110*, 611–646.

Puccini, G. D., Sanchez-Vives, M. V., & Compte, A. (2007). Integrated mechanisms of anticipation and rate-of-change computations in cortical circuits. *PLoS Comput. Biol.*, *3*, e82.

Rugg, M. D., Mark, R. E., Walla, P., Schloerscheidt, A. M., Birch, C. S., & Allan, K. (1998). Dissociation of the neural correlates of implicit and explicit memory. *Nature*, *392*, 595–598.

Rugg, M. D., & Yonelinas, A. P. (2003). Human recognition memory: A cognitive neuroscience perspective. *Trends Cogn. Sci.*, *7*, 313–319.

Standing, L. (1973). Learning 10,000 pictures. *Q. J. Exp. Psychol.*, *25*, 207–222.

Xiang, J., & Brown, M. (2004). Neuronal responses related to long-term recognition memory processes in prefrontal cortex. *Neuron*, *42*, 817–829.

Yakovlev, V., Amit, D. J., Romani, S., & Hochstein, S. (2008). Universal memory mechanism for familiarity recognition and identification. *J. Neurosci.*, *28*, 239–248.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *J. Mem. Lang.*, *46*, 441–517.

Received December 11, 2008; accepted May 14, 2009.