*LETTER* Communicated by Gary Cottrell

# An Autoassociative Neural Network Model of Paired-Associate Learning

**Daniel S. Rizzuto**
**Michael J. Kahana**
*Volen Center for Complex Systems, Brandeis University, Waltham, MA 02454, U.S.A.*

**Hebbian heteroassociative learning is inherently asymmetric. Storing a forward association, from item $A$ to item $B$, enables recall of $B$ (given $A$), but does not permit recall of $A$ (given $B$). Recurrent networks can solve this problem by associating $A$ to $B$ and $B$ back to $A$. In these recurrent networks, the forward and backward associations can be differentially weighted to account for asymmetries in recall performance. In the special case of equal strength forward and backward weights, these recurrent networks can be modeled as a single autoassociative network where $A$ and $B$ are two parts of a single, stored pattern. We analyze a general, recurrent neural network model of associative memory and examine its ability to fit a rich set of experimental data on human associative learning. The model fits the data significantly better when the forward and backward storage strengths are highly correlated than when they are less correlated. This network-based analysis of associative learning supports the view that associations between symbolic elements are better conceptualized as a blending of two ideas into a single unit than as separately modifiable forward and backward associations linking representations in memory.**

## 1 Introduction

To account for performance in standard memory tasks, formal mathematical models of human memory typically employ both autoassociative and heteroassociative mechanisms (Brown, Dalloz, & Hulme, 1995; Humphreys, Bain, & Pike, 1989; Metcalfe, 1991; Murdock, 1993, 1997). Autoassociative information supports the processes of recognition and pattern completion (Metcalfe, 1991; Weber & Murdock, 1989), whereas heteroassociative information supports the processes of paired-associate learning and sequence generation (Chance & Kahana, 1997; Murdock, 1993). The mathematics of matrix memories (Anderson, 1972), often coupled with nonlinear retrieval dynamics (Buhmann, Divko, & Schulten, 1989; Carpenter & Grossberg, 1993; Hopfield, 1982), provides a mechanistic foundation for these models of human associative memory.

This article presents an attractor neural network model of paired-associate learning and uses a model-based analysis of experimental data to

 

shed light on some basic unresolved questions concerning the nature of associations in human memory. The paired-associate learning task is one of the standard assays of human episodic memory. Typically subjects are presented with randomly paired items (e.g., words, letter strings, pictures) and asked to remember each *A*-*B* pair for a subsequent memory test. At test, the *A* items are presented as cues, and subjects attempt to recall the appropriate *B* items.

Attractor network models and classic linear associators provide simple accounts of associative learning. Storing an autoassociation enables pattern completion, and storing a heteroassociation enables recall of paired-stimulus representations. However, if the representations of the to-be-learned items (*A* and *B*) are combined into a single composite representation (by summation or concatenation), it is possible to use an autoassociative network to accomplish heteroassociation. In a simple matrix model, where the representation of the *A*-*B* pair is given by the sum of the constituent item vectors ($\mathbf{a} + \mathbf{b}$), the storage equation for an autoassociative model would be $W_t = W_{t-1} + (\mathbf{a} + \mathbf{b})(\mathbf{a} + \mathbf{b})^T = W_{t-1} + \mathbf{a}\mathbf{a}^T + \mathbf{a}\mathbf{b}^T + \mathbf{b}\mathbf{a}^T + \mathbf{b}\mathbf{b}^T$. Alternatively, one could construct a recurrent heteroassociative matrix memory in which *A* is associated with *B* and *B* is associated back to *A*. In this model, the forward association would be stored in the matrix $W_t^{A\to B} = W_{t-1}^{A\to B} + \mathbf{b}\mathbf{a}^T$, and the backward association would be stored in the matrix $W_t^{B\to A} = W_{t-1}^{B\to A} + \mathbf{a}\mathbf{b}^T$. Although both versions support heteroassociative recall, the autoassociative version assumes symmetric forward and backward associations, whereas the heteroassociative version allows for asymmetric, and possibly independent, forward and backward associations. This distinction between symmetric and asymmetric associations has a long history in the experimental psychology of human memory and learning.

## 2 Associative Symmetry Versus Independent Associations

According to the classic associationist view, the strength of an association is sensitive to the temporal order of encoding (Ebbinghaus, 1885/1913; Robinson, 1932). If *A* and *B* are encoded successively, the forward association, $A \rightarrow B$, is hypothesized to be stronger than the backward association, $B \rightarrow A$. The strengths of forward and backward associations are further hypothesized to be independent (Wolford, 1971). We refer to the strong version of this view as the independent associations hypothesis (IAH). In contrast to this position, representatives of Gestalt psychology (Asch & Ebenholtz, 1962; Köhler, 1947) viewed symbolic associations as composite representations, incorporating elements of each to-be-learned item into a new entity. We refer to this view as the associative symmetry hypothesis (ASH). According to this position, the strengths of forward and backward associations are approximately equal and highly correlated.

In support of the associative symmetry view, early studies of paired-associate learning found approximate equivalence between forward and

backward recall probabilities (see Ekstrand, 1966, for a review). That is, order of study did not seem to influence the strength of forward and backward associations. The equivalence of forward and backward recall was particularly striking when highly imagable words were used as stimuli (Murdock, 1962, 1965, 1966). Data from these experiments showed nearly identical forward and backward recall across variations in presentation rate, serial position, delay between study and test, list length, and number of repetitions of the study pairs. More recent studies have shown that the classic symmetry results are readily replicated (see Kahana, 2001, for a review). This surprising result, supporting gestalt and cognitive ideas, inspired significant debate for nearly two decades. The problem was not resolved, and it resurfaced when mathematical models of associative memory appeared on the scene (Murdock, 1985; Pike, 1984).

Although retrieval in paired associates is approximately symmetric (with respect to order of study), retrieval in free recall and serial recall shows marked asymmetries. For example, when subjects are asked to name the letter that precedes or follows a probe letter in the alphabet, backward retrieval is typically 40 to 60% slower than forward retrieval (Klahr, Chase, & Lovelace, 1983; Scharroo, Leeuwenberg, Stalmeier, & Vos, 1994). In the free recall task, subjects are presented with a series of items and asked to recall them in any order. Analysis of output order reveals that forward transitions are significantly more frequent than backward transitions (Kahana, 1996). This is true of immediate, delayed, and continuous-distractor free recall (Howard & Kahana, 1999).

Although many studies have found symmetry between forward and backward recall, there are still conditions under which symmetric retrieval is violated. For example, in studies where pairs of items are chosen from different linguistic classes, strong asymmetry is often observed. Lockhart (1969) showed that cued recall of noun-adjective pairs was superior when cued with the noun than when cued with the adjective, independent of the order of study. However, when subjects study noun-noun pairs, associative symmetry is observed. Wolford (1971) examined both digit-word and word-digit pairs and found an advantage when using words as the retrieval cue, irrespective of the order of presentation.

Kahana (2001) presented an analysis of the implications of associative symmetry for linear matrix and convolution models of human associative memory. In his treatment, he showed that models assuming symmetry (Metcalfe, 1991; Murdock, 1997) can still account for empirical asymmetries in overall forward and backward recall performance. Consider, for example, items with different numbers of preexperimental associates. If item $A$ has many more preexperimental associates than item $B$, then probing memory with item $A$ will result in significant associative interference and impair subjects' performance. This will occur even if the $A$-$B$ association is stored symmetrically.

Although models that assume symmetry can account for findings of retrieval asymmetries, models that assume independent associations can also

account for the equivalence of forward and backward recall performance. In this case, if associations are formed independently but with equal strength on average, symmetric retrieval probabilities would be realized. It is thus unlikely that forward and backward recall probabilities will discriminate among competing theories.

To address the question of the dependence or independence of forward and backward associations, Kahana (2001) employed the method of successive tests, which involves successively testing the stored associations in both the forward and reverse directions. With this method, one can directly estimate the correlation between forward and backward recall of a given studied pair. Kahana reported experimental evidence consistent with the ASH.

## 3 A Neural Network Model of Paired-Associate Learning

We developed an autoassociative neural network model to simulate forward and backward associative recall and to model data on the correlation between successive recall tests by human subjects (Metcalfe, 1991; Metcalfe, Cottrell, & Mencl, 1993). Although autoassociation is generally used to encode a single representation, if the representation being stored is the combination of two items, it is possible to store both autoassociative and heteroassociative information within a single memory matrix. Assuming that the representations of the $A$ and $B$ items are concatenated, the general form of the storage equation for a list of $L$ pairs is given by

$$W = \sum_{\nu=1}^{L} (\mathbf{a}^{\nu} \oplus \mathbf{b}^{\nu})(\mathbf{a}^{\nu} \oplus \mathbf{b}^{\nu})^{T}, \tag{3.1}$$

where $\mathbf{a}^{\nu}$ and $\mathbf{b}^{\nu}$ are binary ($\pm 1$), $N$-element vectors representing the items to be associated, $\mathbf{a}^{\nu} \oplus \mathbf{b}^{\nu}$ denotes the concatenation of $\mathbf{a}^{\nu}$ and $\mathbf{b}^{\nu}$, and $W$ is the $2N \times 2N$ weight matrix. This outer product produces a memory matrix with four quadrants, as shown in Figure 1. Quadrants 1 and 3 of the matrix contain autoassociative information, and quadrants 2 and 4 contain heteroassociative information.

We use a probabilistic encoding[1] algorithm that enables the network to account for the effect of repetition on performance. Each weight in the matrix has some probability of being stored correctly, depending on the

[1] In an analysis of different learning rules for distributed memory models, Murdock (1989) found that probabilistic encoding of the individual "features" of item vectors provided a good fit to data on paired-associate learning. In this framework, each feature is encoded with some probability on each presentation of the item. This results in an increase in the number of features encoded with each presentation. For more recent models using probabilistic learning in autoassociative networks, see Amit and Brunel (1995), Amit and Fusi (1992), and Brunel, Carusi, and Fusi (1998).

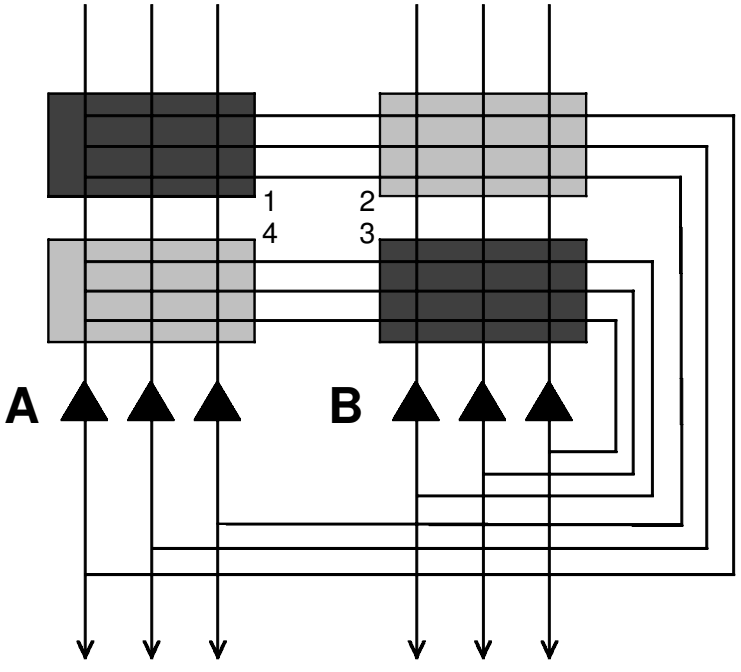


Figure 1: Graphical representation of the autoassociative network architecture used to simulate the experimental paradigm. Light rectangles indicate heteroassociative quadrants of the weight matrix, and the darker rectangles indicate autoassociative quadrants. Triangles represent the nodes in the network.

quadrant of the matrix in which it resides and the number of repetitions. The heteroassociative quadrants of the matrix drive the associative recall process, and we introduce two random variables, $\gamma_f$ and $\gamma_b$, which control learning in the forward and the backward directions, respectively.

For a pair in the list, the rule for storing each heteroassociative weight in the quadrant that mediates forward recall (quadrant 2 in Figure 1) is given by

$$\Delta W_{ij} = \begin{cases} s_i^{\nu} s_j^{\nu} & \text{with probability} \quad \gamma_f \\ 0 & \text{with probability} \quad 1 - \gamma_f, \end{cases} \tag{3.2}$$

where $\mathbf{s}^{\nu} = (\mathbf{a}^{\nu} \oplus \mathbf{b}^{\nu})$ and $\gamma_f \sim \mathcal{N}(\mu, \sigma)$. Similarly, the probability of storing each heteroassociative weight that mediates backward recall is given by $\gamma_b \sim \mathcal{N}(\mu, \sigma)$. The parameter $\mu$ represents the mean probability of encoding, while the parameter $\sigma$ represents the variability in encoding across pairs.

If $\gamma_f$ and $\gamma_b$ are perfectly correlated, the model implements the ASH. If $\gamma_f$ and $\gamma_b$ are independent of one another, the model implements the IAH. Rather than pitting the two hypotheses against each other, we can allow the model to determine the correlation between $\gamma_f$ and $\gamma_b$, denoted $\rho(\gamma_f, \gamma_b)$, that best fits the data. The power of this parameter lies in its ability to change the behavior of the model from that approximating the IAH when $\rho$ approaches zero, to that approximating the ASH when $\rho$ approaches unity. For the correlation parameter to be meaningful, the values of $\gamma_f$ and $\gamma_b$ must vary across pairs. This reflects the empirical finding that some pairs are easier to learn than others. The correlation represents the degree to which variation in encoding is at the level of pairs within a list versus the level of forward and backward associations within a pair. Despite the variation in $\rho$, the mean level of encoding is the same for both the forward and backward associations. This means that average forward recall and backward recall will be equivalent even when $\rho$ approaches zero.

Thus, for each *A*-*B* pair, the learning of the forward and backward association is determined by the distributions of $\gamma_f$ and $\gamma_b$, as well as the correlation, $\rho$, between these two parameters. Because associative recall is driven by the heteroassociative quadrants of the matrix, we make the simplifying assumption of no variability in autoassociative encoding across pairs. Thus, the weights in quadrants 1 and 3 of the matrix are stored equally well for every item, with the probability of storage equal to $\mu$.

Sampling the learning probabilities $\gamma_f$ and $\gamma_b$ from a normal distribution of mean $\mu$ and variance $\sigma^2$ occasionally produces probabilities that fall outside the range of 0 through 1. These probabilities are meaningless, and when this occurs, both variables are redrawn from the distribution, thus producing a truncated normal distribution. Using this method produces an effective mean and variance that are somewhat different from the parameters used to generate the distribution.

Retrieval in this model follows Hopfield (1982). Recall of pair $\nu$ proceeds asynchronously, with a random node being updated on each iteration, $t$, and its activation given by

$$s_i^{\nu}(t+1) = \operatorname{sgn}\left(\sum_j W_{ij}\, s_j^{\nu}(t)\right). \tag{3.3}$$

The initial state of the network, $\mathbf{s}^{\nu}(0)$, is equal to $(\mathbf{a}^{\nu} \oplus \mathbf{k})$ for forward recall and $(\mathbf{k} \oplus \mathbf{b}^{\nu})$ for backward recall, where $\mathbf{k}$ is an $N$-dimensional, random, binary ($\pm 1$) vector, uncorrelated from trial to trial. The state of the cue vector ($\mathbf{a}^{\nu}$, in the case of a forward test) is clamped throughout retrieval.

After each iteration, the state of the network is compared to the target item. If the similarity between the network state and the target, as measured by the cosine of the angle between the two vectors, is greater than a constant criterion, $\theta$, then the target is said to be recovered. If not, the process

repeats until a maximum number of iterations has been reached, denoted $I_{\max}$. At this time, if the network has not reached the criterion level of item recovery, that item is considered nonrecallable. In the work presented here, the parameters $\theta$, $I_{\max}$, and $N$ were fixed to 0.99, 800, and 70, respectively; changing them had very little effect on the results of the simulations.

## 4 Applying the Model to a Paired-Associate Learning Data Set

We applied our model to data gathered from a paired-associate learning experiment reported in Kahana (2001). First, we briefly describe the experiment, and then we report model fits to the data.

The experimental procedure is illustrated in Figure 2. Subjects studied a list of 12 unique word pairs. Each pair was presented visually for 2 seconds, and subjects were instructed to read the words aloud from left to right to ensure that they processed the two words in temporal succession. To assess learning, equal numbers of word pairs were presented either one, three, or five times in each list. The order of presentation was random, subject to the constraint that identical pairs were never repeated successively. After studying the list of word pairs, subjects performed a distractor task (pattern matching) aimed at minimizing the role of recency-sensitive retrieval processes (Howard & Kahana, 1999).

Each pair was tested for recall once in each of two successive test phases (test 1 and test 2). In test 1, subjects were tested on each of the studied pairs: half in the forward direction ($A \rightarrow ?$) and half in the backward direction ($? \leftarrow B$). In test 2, half of the pairs that were first tested in the forward direction were tested in the backward direction, and the other half were again tested in the forward direction. The same was true of pairs that were tested in the backward direction on test 1. This produced a $2 \times 2$ factorial of test 1–test 2 possibilities (forward-forward, forward-backward, backward-forward, and backward-backward).

For each of the four combinations of test 1–test 2 possibilities, one can construct a $2 \times 2$ contingency table tallying the outcomes of the first and second tests (success and failure on test 1 crossed with success and failure on test 2; see Figure 3). The marginal probabilities of this contingency table yield the probability of success on test 1 (cell $a$ + cell $c$) and the probability of success on test 2 (cell $a$ + cell $b$). Additionally, one can measure the dependency between the two outcomes using a standard measure of association (e.g., Yule's $Q$, Goodman-Kruskal's $\gamma$ , or $\chi^2$). We adopt Yule's $Q$ because of its extensive use in the study of memory (see Kahana, 1999, for a review) and for its desirable statistical properties (Bishop, Fienberg, & Holland, 1975). This dependency, or correlation, between the outcomes of the two tests provides valuable information that is not contained in the marginal probabilities of recall.

The experimental results and model fits are shown in Tables 1 and 2. Table 1 reports recall probability and test 1–test 2 correlations (Yule's $Q$) as

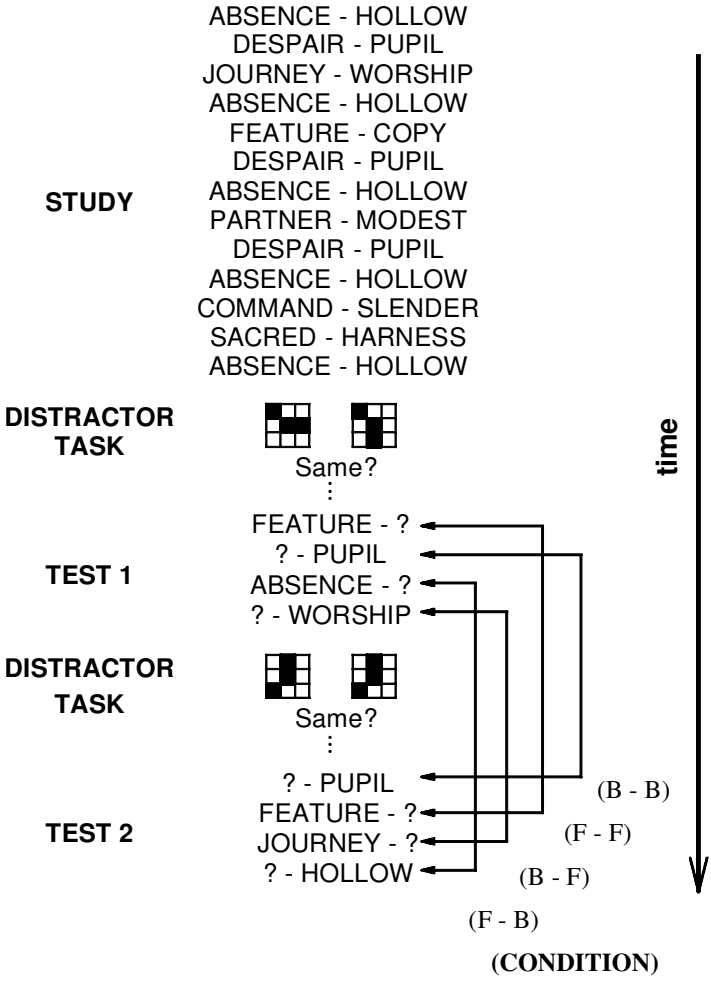


Figure 2: Depiction of the experimental procedure. The study session is followed by a distractor task (pattern matching), which is followed by test 1, another distractor task, and then test 2.

a function of the number of presentations, averaged across subjects. From the behavioral data, it can be seen that there were no significant differences between forward and backward recall probabilities within any presentation level of the experiment. Also, correlations between identical successive tests were very high (near one); correlations between reverse successive tests were also high, but not as high as for the identical successive tests.

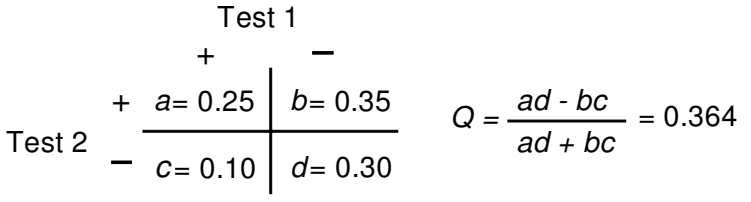


Figure 3: Example of a 2 × 2 contingency table, illustrating the calculation of the Yule's $Q$ measure of association between test 1 and test 2 performance.

Table 1: Observed Accuracy and Correlation Values for Each Presentation Level in the Experiment, Averaged Across Subjects.

| | Probability of Recall | | | | Yules $Q$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Forward | | Backward | | Identical | | Reversed | |
| | Data | Model | Data | Model | Data | Model | Data | Model |
| *1p* | 0.35 | 0.35 | 0.36 | 0.35 | 0.97 | 0.99 | 0.84 | 0.85 |
| *3p* | 0.65 | 0.65 | 0.65 | 0.65 | 0.96 | 0.99 | 0.81 | 0.83 |
| *5p* | 0.75 | 0.75 | 0.73 | 0.75 | 0.96 | 0.99 | 0.84 | 0.82 |

In Table 2 we report the average of the individual subject's contingency tables for the behavioral data and the model.[2] We fit the model to each individual subject's contingency table separately. In doing so, we allowed $\mu$ and $\sigma$ to vary independently for each level of learning ($\mu_1$, $\mu_3$, $\mu_5$, $\sigma_1$, $\sigma_3$, and $\sigma_5$). This quantitatively fits the learning data without constraining the model to a particular learning mechanism. A single $\rho$ is used to correlate forward and backward learning for all presentation levels. For parameter estimation, we minimize the model's root-mean-squared deviation (RMSD) from the behavioral data for each of the 15 subjects who took part in the experiment.[3] We fit the model concurrently to the contingency tables for

[2] Although Tables 1 and 2 are computed from the same underlying behavioral data, computing Yule's $Q$ values from Table 2 will not necessarily produce the values seen in Table 1. The reason for this lies in the fact that when calculating Yule's $Q$, it is customary to add 0.5 to each cell in the contingency table. This is due to the fact that Yule's $Q$ is a ratio and is thus extremely sensitive to cells having values of zero or near zero. Table 2 contains raw contingency tables averaged across subjects without adding 0.5 to each cell. This is appropriate because no derived measure is being calulated. Thus, Table 2 represents the average contingency table and was not used when calculating the average value of Yule's $Q$.

[3] Error minimization was accomplished via a genetic algorithm (Mitchell, 1996). Fifteen populations (one for each subject in the experiment) of 100 random points ("individuals") in parameter space were evolved for many generations until the average fitness for the population did not change from one generation to the next. An individual's fitness was calculated to be the negative of its RMSD compared to the behavioral data for a particular

Table 2: Contingency Tables from the Behavioral Experiment Compared with Simulated Contingency Tables.

| | **Identical Direction** | | | | **Reversed Directions** | | | |
|---|---|---|---|---|---|---|---|---|
| | Data | | Model | | Data | | Model | |
| *1p* | 0.319 | 0.012 | 0.321 | 0.028 | 0.293 | 0.122 | 0.262 | 0.100 |
| | 0.006 | 0.663 | 0.032 | 0.620 | 0.049 | 0.537 | 0.097 | 0.542 |
| *3p* | 0.583 | 0.012 | 0.625 | 0.026 | 0.609 | 0.074 | 0.575 | 0.084 |
| | 0.037 | 0.368 | 0.027 | 0.323 | 0.043 | 0.273 | 0.088 | 0.254 |
| *5p* | 0.729 | 0.018 | 0.723 | 0.025 | 0.681 | 0.086 | 0.665 | 0.083 |
| | 0.012 | 0.241 | 0.019 | 0.233 | 0.037 | 0.196 | 0.078 | 0.173 |

the identical and reversed successive tests for all three levels of learning, thereby fitting 24 data points with seven free parameters. Figure 4 plots the best-fitting model parameters with 95% confidence intervals calculated across subjects. By fitting each (normalized) quadrant of the contingency tables from the experimental data, the model was able simultaneously to fit correlations and accuracy on tests 1 and 2 (RMSD = 0.07).

These model fits to the experimental data are interesting in several respects. As expected, we see that the mean level of encoding increases with learning (from $\mu_1$ to $\mu_3$ to $\mu_5$). We also see that significant variability in learning across pairs ($\sigma$) is required to fit the experimental data. Although this is consistent with psychological studies of the role of variability in learning (Hintzman & Hartry, 1990), this important facet of human performance has rarely been incorporated into neural network or even formal mathematical models of learning and memory. Finally, and most important, the fact that the best-fitting values of $\rho$ are very high (near one) tells us that very strong correlations between forward and backward storage are necessary to fit the human behavioral data.

Once the best parameter set for each subject was found, we completed a comprehensive search of the local parameter space to look at the effects of correlations and variability in learning on the fitness landscape. In Figure 5 we plot model fitness as a function of $\rho$ and $\sigma_3$; each of the remaining parameters was kept fixed. Fitness was calculated using only the contingency tables for three presentations.

Underneath each of the simulated data points in the fitness landscape, we plot the $p$-value representing the significance of the difference between that parameter set and the best-fitting parameter set. Lower $p$-values reflect

subject. Each individual was run for 300 trials of 12-pair lists in order to generate reliable values. At the end of every generation, each of the 50 least-fit individuals were completely regenerated from 2 of the best-fitting individuals by randomly drawing their new parameter values from each of their parents. Those 50 individuals with the best fitness were mutated by a single, gaussian parameter change.

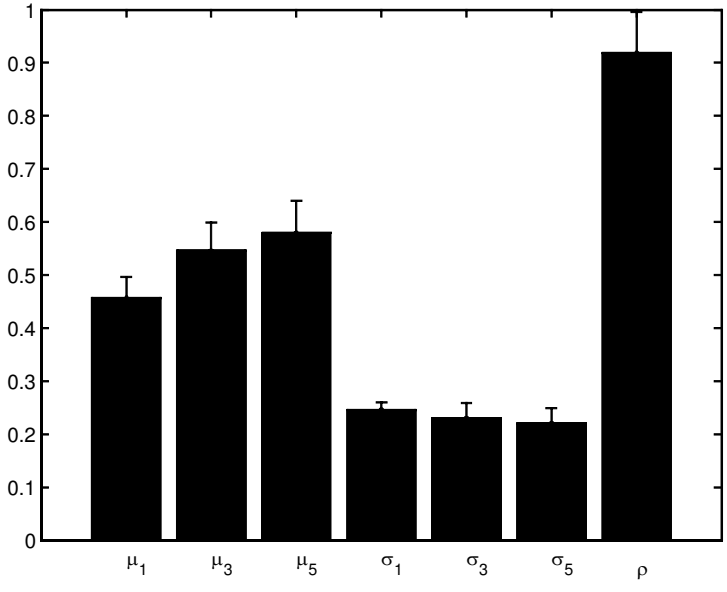


Figure 4: Best-fitting parameter values averaged across subjects. Error bars represent 95% confidence intervals.

an inability of that parameter set to fit the behavioral data as well as the best-fitting parameter set. Figure 5 shows that the best-fitting portion of the parameter space ($\rho \geq 0.8$ and $0.15 \leq \sigma_3 \leq 0.25$) fits significantly better than the rest of the space.

We used the same simulations to assess correlations between forward and backward recall. Figure 6 plots the correlation between simulated successive tests in opposite directions as a function of both $\rho$ and $\sigma_3$. In this figure, it is shown that the only way of producing high correlations in the retrieval of forward and backward associations is to have extremely high correlations in the learning of those associations. This figure also shows the monotonic relationship between the correlations in learning the forward-backward associations and the correlation in recall between successive tests in opposite directions. The correlation between successive tests in the identical direction is always one.

It may be seen that Yule's $Q$ decreases as encoding variability ($\sigma$) across items decreases. This occurs when $\sigma$ is small compared to the binomial variability introduced during probabilistic encoding. Although there is no difference between the probabilities of encoding the forward and the backward associations when $\sigma = 0$, there are random fluctuations in the actual number of weights stored for each association (e.g., if there are 1000 weights in each quadrant of the matrix and $\gamma_f = \gamma_b = 0.5$, one trial might store 512

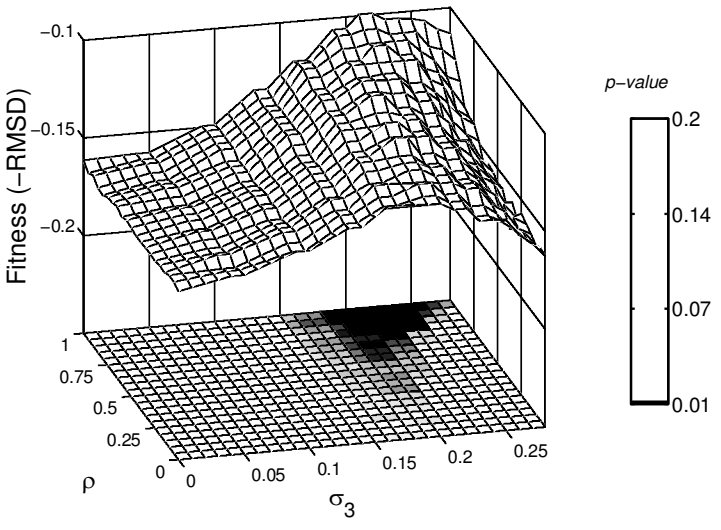


Figure 5: Model fitness (negative RMSD) plotted on the $z$-axis as a function of $\rho$ and $\sigma_3$. Below the fitness function, we plot the $p$-values corresponding to the statistical significance of the difference between the best-fitting parameter set and the rest of the parameter space.

weights in the forward weight matrix while storing only 487 weights in the backward weight matrix). As $\sigma \to 0$, encoding probabilities are very similar for all pairs, and the uncorrelated noise associated with binomial variability leads to much lower correlations in recall.

## 5 Output Encoding

In the model just described, we assumed that the test trials did not affect information stored in memory. Although this assumption is common in neural network models, behavioral studies suggest that test trials do contribute to learning (Humphreys & Bowyer, 1980). If subjects are reencoding the pair during correct responses to the first test, this could increase the probability of responding correctly to the second test and artifactually increase the observed correlation.[4]

To explore this possibility, we augmented our basic model by introducing an output encoding parameter. This parameter, $\phi$, regulates the reencoding

[4] Although it is possible that subjects will encode some representation on incorrect trials as well, subjects rarely make errors of commission in these experiments. They usually respond with "Pass" instead.

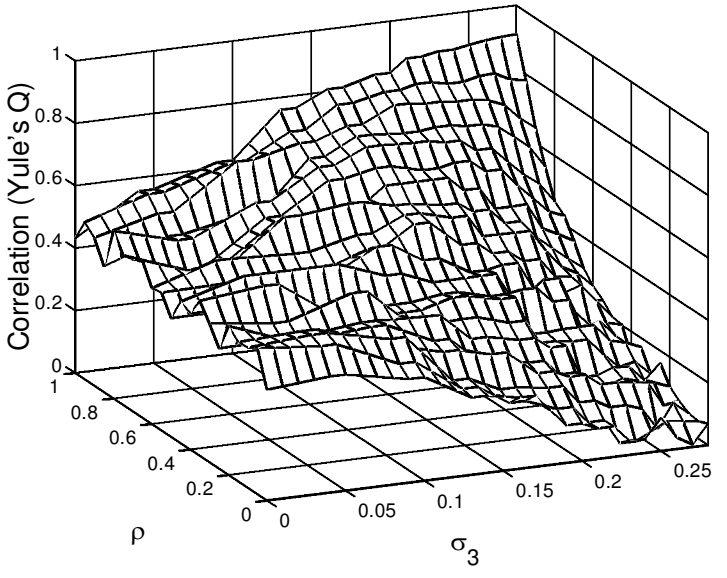


Figure 6: Simulated correlation (Yule's $Q$) between forward and backward recall as a function of $\sigma_3$ and $\rho$.

of the original association on correct trials. When a test item is correctly recalled, each weight in all of four matrix quadrants given by equation 3.1 is probabilistically reencoded as

$$\Delta W_{ij} = \begin{cases} s_i^{\nu} s_j^{\nu} & \text{with probability} \quad \phi \\ 0 & \text{with probability} \quad 1 - \phi. \end{cases} \tag{5.1}$$

This implementation of output encoding does not include variability, and the probability of storing a weight is independent of the quadrant it resides in. The fact that we are reencoding both the forward and the backward associations after a correct recall in either direction should increase the simulated correlation between forward and backward recall.

In order to examine the contribution of output encoding to the behavioral predictions of our model, we again fit all eight parameters of the model to the experimental data for each subject. Figure 7 plots the best-fitting values of $\mu$, $\sigma$, $\rho$, and $\phi$ averaged across individual subjects. As before, $\mu$ increases monotonically with the number of presentations of the studied pair, and $\sigma$ stays relatively constant across the three learning levels, corroborating the simulations without output encoding. These simulations differ, however, with respect to $\rho$, which settles into a lower value with output encoding than without (compare Figure 7 with Figure 4). This difference was not

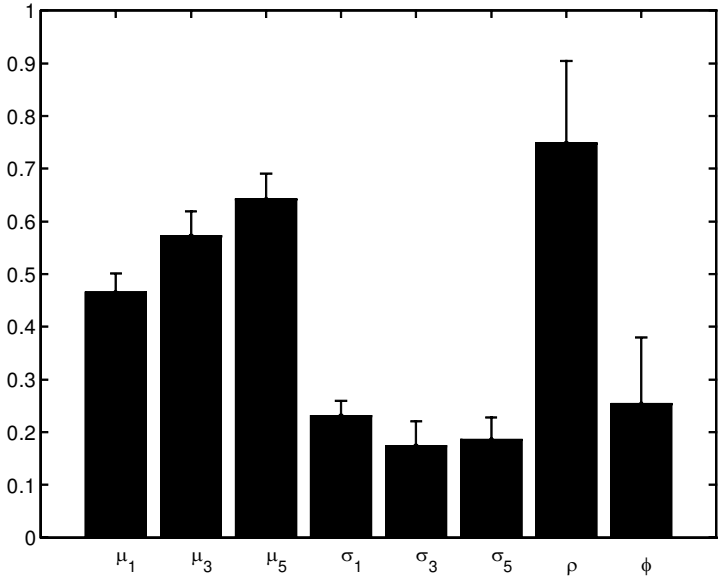


Figure 7: Best-fitting parameter values for the model that includes output encoding. Error bars represent 95% confidence intervals.

statistically significant ($t(28) < 1$, n.s.). Similarly, the augmented model does not fit the data significantly better than the model without output encoding (RMSD = 0.06; $t(28) < 1$, n.s.). It should be noted however, that the best-fitting value of $\rho$ was significantly less than one ($t(14) = 3.15$, $p < 0.01$), suggesting that although the ASH provides a much better description of behavior than the IAH, the true state of the world may fall shy of perfect symmetry.

## 6 Conclusion

The finding of very strong correlations between forward and backward recall of paired associates in human memory supports the view that both forward and backward associations are products of a single mechanism. The strength of our model lies in its ability to modulate its behavior accurately and flexibly between two extreme views of associative learning: ASH and the IAH, via the correlational parameter, $\rho$. The fact that the average best-fitting parameter set (without output encoding) included a value of $\rho$ close to one is strong evidence for symmetrical learning of associations in humans.

These results lend support to models that employ inherently symmetric associative mechanisms, like the holographic models of Murdock (1982, 1997) and Metcalfe (1991; Metcalfe Eich, 1982). Although the neural plau-

sibility of convolution-based models has been called into question (Pike, 1984, but see Murdock, 1985), they have been successfully applied to a wide range of experimental data.

We have examined the possibility that output encoding during the first test is affecting the results of the second test, a hypothesis first explored by Humphreys and Bowyer (1980) in the context of successive recognition-recall tests. Although our model can fit the human behavioral data marginally better when it includes output encoding than when it does not, this difference is not statistically significant. Most important, the best-fitting value of $\rho$ does not change significantly with the inclusion of an output encoding mechanism.

This model also addresses the effects of variability in the learning of associations. The best-fitting parameter set at all levels of learning contained significant variability in the distribution of learning probabilities. Clearly there are differences in encoding at the level of individual word pairs; some items are easier to learn than others, and fluctuations in attention can provide an additional source of variability. The current results support the idea that variability in learning is an important factor that must be considered in order to model the human learning process accurately.

The idea that an autoassociative mechanism may underlie both item retrieval (redintegration) and associative retrieval (cued recall) is not unreasonable given what we know about hippocampal physiology. Recurrent collaterals in the CA3 region of the hippocampus may provide a neurophysiological basis for autoassociative memory (Treves & Rolls, 1994). Each CA3 cell receives input from mossy fibers arising in the dentate gyrus and direct perforant path inputs from entorhinal cortex (EC). Yet by far the majority of the connections come from other CA3 cells (Ishizuka, Cowan, & Amaral, 1995). If incoming sensory inputs arriving from association cortex via EC contain a composite representation of both *A* and *B* items, this strong recurrent connectivity could provide a symmetric memory storage mechanism. Since synaptic modifiability within this system occurs on such a short timescale (Hanse & Gustafsson, 1994), it would be well suited to forming cohesive, episodic representations.

Like humans, rats can also learn associations in the forward direction and exhibit transfer of the same association tested in the backward direction. Bunsey and Eichenbaum (1996) trained rats to associate a stimulus odor with one of two response odors in an olfactory learning paradigm. They then assessed the rats' ability to choose the stimulus odor when presented with its paired response. Although transfer was not perfect, it was substantial ($> 50\%$). Following hippocampal lesions, these animals still show significant retention of the forward association but exhibit no transfer when tested in the reverse order. On the basis of these results, they suggest that there may be two mechanisms involved in the formation of associations. A declarative mechanism, mediated by the hippocampal formation, allows for flexible usage of associative relations, thus supporting both forward and

backward associative learning. A second procedural mechanism supports the acquisition of stimulus-stimulus associations but not the inferential processing needed to retrieve backward associations. This second mechanism does not depend on an intact hippocampal formation.

A convergence of anatomical, behavioral, and physiological evidence points to the hippocampus as an integrator of multimodal sensory information, crucial to memory processing. The fact that it provides the necessary infrastructure to support an autoassociative memory system also suggests the intriguing hypothesis that this type of architecture may indeed underly both auto- and heteroassociative memory in humans.

In summary, our autoassociative network model of heteroassociative memory implements a stochastic learning algorithm acting at the level of individual weights and allows us quantitatively to fit human accuracy data as well as the correlations between successive tests. Fitting these experimental data highlights the importance of variability in the learning process. In addition, data on the correlations between successive forward and backward recall tests support the notion that an autoassociative mechanism may underlie at least some forms of heteroassociative learning.

## Acknowledgments

We acknowledge support from National Institutes of Health research grants MH55687 and AG15852. Correspondence concerning this article should be addressed to Michael Kahana, Volen Center for Complex Systems, MS 013, Brandeis University, Waltham, MA 02254-9110. Electronic mail may be sent to kahana@brandeis.edu.

## References

Amit, D. J., & Brunel, N. (1995). Learning internal representations in an attractor neural network with analogue neurons. *Network: Computation in Neural Systems, 6*, 359–388.

Amit, D. J., & Fusi, S. (1992). Constraints on learning in dynamic synapses. *Network: Computation in Neural Systems, 3*, 443–464.

Anderson, J. A. (1972). A simple neural network generating an interactive memory. *Mathematical Biosciences, 14*, 197–220.

Asch, S. E., & Ebenholtz, S. M. (1962). The principle of associative symmetry. *Proceedings of the American Philosophical Society, 106*, 135–163.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

Brown, G. D. A., Dalloz, P., & Hulme, C. (1995). Mathematical and connectionist models of human memory: A comparison. *Memory, 3*, 113–145.

Brunel, N., Carusi, F., & Fusi, S. (1998). Slow stochastic Hebbian learning of classes of stimuli in a recurrent neural network. *Network: Computation in Neural Systems, 9*, 123–152.

Buhmann, J., Divko, R., & Schulten, K. (1989). Associative memory with high information content. *Physical Review A, 39*, 2689–2692.

Bunsey, M., & Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature, 379*, 255–257.

Carpenter, G., & Grossberg, S. (1993). Normal and amnesic learning, recognition and memory by a neural model of cortico-hippocampal interactions. *Trends in Neurosciences, 16*, 131–137.

Chance, F. S., & Kahana, M. J. (1997). Testing the role of associative interference and compound cues in sequence memory. In J. Bower (Ed.), *The neurobiology of computation: Proceedings of the Annual Computational Neuroscience Meeting*. Norwell, MA: Kluwer.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. New York: Teachers College, Columbia University.

Ekstrand, B. R. (1966). Backward associations. *Psychological Bulletin, 65*, 50–64.

Hanse, E., & Gustafsson, B. (1994). Onset and stabilization of NMDA receptor-dependent hippocampal long-term potentiation. *Neuroscience Research, 20*, 15–25.

Hintzman, D., & Hartry, A. L. (1990). Item effects in recognition and fragment completion: Contingency relations vary for different subsets of words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 965–969.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 84*, 8429–8433.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*, 923–941.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review, 96*, 208–233.

Humphreys, M. S., & Bowyer, P. A. (1980). Sequential testing effects and the relation between recognition and recognition failure. *Memory and Cognition, 8*, 271–277.

Ishizuka, N., Cowan, W. M., & Amaral, D. G. (1995). A quantitative analysis of the dendritic organization of pyramidal cells in the rat hippocampus. *Journal of Comparative Neurology, 362*, 17–45.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition, 24*, 103–109.

Kahana, M. J. (1999). Contingency analyses of memory. In E. Tulving (Ed.), *Oxford handbook of human memory* (pp. 323–384). New York: Oxford Press.

Kahana, M. J. (2001). Associative symmetry and memory theory. Submitted.

Klahr, D., Chase, W. G., & Lovelace, E. A. (1983). Structure and process in alphabetic retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*, 462–477.

Köhler, W. (1947). *Gestalt psychology*. New York: Liveright.

Lockhart, R. S. (1969) . Retrieval asymmetry in the recall of adjectives and nouns. *Journal of Experimental Psychology, 79*, 12–17.

Metcalfe, J. (1991). Recognition failure and the composite memory trace in CHARM. *Psychological Review, 98*, 529–553.

Metcalfe, J., Cottrell, G. W., & Mencl, W. E. (1993). Cognitive binding: A computational-modeling analysis of a distinction between implicit and explicit memory. *Journal of Cognitive Neuroscience, 4*, 289–298.

Metcalfe Eich, J. (1982) . A composite holographic associative recall model. *Psychological Review, 89*, 627–661.

Mitchell, M. (1996). *An introduction to genetic algorithms*. Cambridge, MA: MIT Press.

Murdock, B. B. (1962). Direction of recall in short term memory. *Journal of Verbal Learning and Verbal Behavior, 1*, 119–124.

Murdock, B. B. (1965). Associative symmetry and dichotic presentation. *Journal of Verbal Learning and Verbal Behavior, 4*, 222–226.

Murdock, B. B. (1966). Forward and backward associations in paired associates. *Journal of Experimental Psychology, 71*, 732–737.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*, 609–626.

Murdock, B. B. (1985). Convolution and matrix systems: A reply to Pike. *Psychological Review, 92*, 130–132.

Murdock, B. B. (1989). Learning in a distributed memory model. In C. Izawa (Ed.), *Current issues in cognitive processes: The Floweree Symposium on Cognition* (p. 69–106). Hillsdale, NJ: Erlbaum.

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review, 100*, 183–203.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review, 104*, 839–862.

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review, 91*, 281–294.

Robinson, E. S. (1932). *Association theory today: An essay in systematic psychology*. New York: Century Co.

Scharroo, J., Leeuwenberg, E., Stalmeier, P. F. M., & Vos, P. G. (1994). Alphabetic search: Comment on Klahr, Chase, and Lovelace (1983). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 234–244.

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

Weber, E. U., & Murdock, B. B. (1989). Priming in a distributed memory system: Implications for models of implicit memory. In S. Lewandowsky, J. Dunn, & K. Kirsner (Eds.), *Implicit memory: Theoretical issues* (pp. 87–89). Hillsdale, NJ: Erlbaum.

Wolford, G. (1971). Function of distinct associations for paired-associate performance. *Psychological Review, 78*, 303–313.

Received April 22, 1999; accepted November 2, 2000.