Asymmetric Weights and Retrieval Practice in an Autoassociative Neural Network Model of Paired-Associate Learning

Sneha Aenugu
saenugu@umass.edu
David E. Huber
dehuber@umass.edu
Department of Psychological and Brain Sciences, University of Massachusetts, Amherst, MA 01003, U.S.A.

Rizzuto and Kahana (2001) applied an autoassociative Hopfield network to a paired-associate word learning experiment in which (1) participants studied word pairs (e.g., ABSENCE-HOLLOW), (2) were tested in one direction (ABSENCE-?) on a first test, and (3) were tested in the same direction again or in the reverse direction (?-HOLLOW) on a second test. The model contained a correlation parameter to capture the dependence between forward versus backward learning between the two words of a word pair, revealing correlation values close to 1.0 for all participants, consistent with neural network models that use the same weight for communication in both directions between nodes. We addressed several limitations of the model simulations and proposed two new models incorporating retrieval practice learning (e.g., the effect of the first test on the second) that fit the accuracy data more effectively, revealing substantially lower correlation values (average of .45 across participants, with zero correlation for some participants). In addition, we analyzed recall latencies, finding that second test recall was faster in the same direction after a correct first test. Only a model with stochastic retrieval practice learning predicted this effect. In conclusion, recall accuracy and recall latency suggest asymmetric learning, particularly in light of retrieval practice effects.

To promote stability, autoassociative Hopfield (1982) networks are typically implemented with symmetric learning, using the same weight value in both directions (MacKay, 2003). However, several prominent synaptic learning rules are asymmetric, based on timing (Bi & Poo, 1998) or activation differences (Bienenstock, Cooper, & Munro, 1982). Relatedly, in the study of human memory, it is unclear whether a learned association in one direction (stork -> baby) necessarily produces learning in the opposite direction (baby -> stork). To address this question, Kahana (2002) ran a paired-associate learning experiment in which participants studied unrelated word

https://doi.org/10.1162/neco_a_01444

pairs (one word presented on the left, paired with a word presented on the right), followed by a cued recall test of all pairs (half in the forward direction and half in the backward direction), and then a second test of all pairs, with half of the second tests in the same direction as the first test while the other half were in the reverse direction.

Kahana (2002) found similar recall accuracy in the forward (left word -> ?) and backward (? <- right word) directions, suggesting symmetry. However, average results can be misleading, owing to random subject and word-pair effects that can artificially produce symmetry. For instance, a participant who does well on the first test is likely to do well on the second test because he or she has better memory in general. Similarly, a word pair that produces a correct response on the first test is likely to produce a correct response on the second test if that word pair was studied with high attention. To de-confound the analysis from these random effects, Rizzuto and Kahana (2001) fit a Hopfield network to each participant separately, including a word-pair variability parameter (σ) to factor out word-pair effects. The model was applied to the joint probability data of yes/no accuracy on the first test crossed with yes/no accuracy on the second test.

Rizzuto and Kahana's (2001) model used the same 140 nodes for all word pairs, with random patterns of 70 $+1/-1$ values representing each word. Seventy nodes were used to represent all words that appeared on the left side of the screen during initial study, while the remaining 70 nodes were used to represent all words that appeared on the right side. Every pair of nodes was connected with both a forward and a backward weight, and whether weight updating occurred was stochastic for each of the two directed weights (if updated, the learning rate was 1.0). Thus, when learning a word pair, some connections were updated in both directions, some were updated in neither direction, and some were updated in only one direction.

The key mechanism for assessing symmetry was whether the weight updating probability was the same in the forward and backward directions for a given word pair. For pairs of nodes belonging to the same word, the weight-updating probability was fixed to the same value (μ) in both directions (symmetric updating within a word). For pairs of nodes belonging to different words, the forward and backward weight update probabilities were determined separately for each word pair by taking a sample from a bivariate normal distribution for the word pair, thus determining if the word pair was symmetric (i.e., whether the same weight updating probability was used in both directions). The bivariate normal distribution contained three free parameters: the mean (μ) determined the average weight updating probabilities in both directions, the standard deviation (σ) determined weight updating variability in both directions, and the correlation (ρ) determined the dependency between weight updating in one direction versus the other. Thus, if the correlation was small, the association between two words of a word pair might be asymmetric (see also appendix A).

Practice tests are a powerful form of learning (Abbott, 1909; Roediger & Butler, 2011) and Rizzuto and Kahana applied a version of the model with an additional parameter for the probability of learning from a successful test. However, this version did not fit any better than the model without retrieval practice, and their conclusions were based on the model without retrieval practice, which produced best-fitting ρ values close to 1.0 for all participants, supporting the conclusion that learning is symmetric. Recently, this symmetry assumption has been included in formal models of paired-associate learning (Cox & Criss, 2020; Polyn, Norman, & Kahana, 2009; Popov & Reder, 2020).

We identified several limitations of the Rizzuto and Kahana simulations, casting doubt on the conclusion that newly learned associations between words of a word pair are symmetric. For one, we found that the model architecture produced implausible interference effects; learning of left words produced interference for other left words but not right words, and yet screen position of words rarely matters in list-learning experiments (Pezdek, Roman, & Sobolik, 1986). We remedied this problem by randomly assigning for each word pair 70 nodes to the left word and the remaining 70 to the right word, with this assignment differing across word pairs. This contrasts with the Rizzuto and Kahana approach, which used the same assignment of nodes to left-words versus right-words for all word pairs. Thus, in our approach, the forward/backward weights connecting two nodes were updated using the same update probability when studying a word for which the two nodes were both randomly assigned to that word (e.g., both assigned to the left-word). However, for other word pairs, the same two nodes might be assigned to different words (one being assigned to the left-word while the other assigned to the right-word), and upon studying such a word pair, the forward weight might be updated with a probability different from the backward weight. All other aspects of the simulation were the same as the original study, except as noted.

To highlight the effect of retrieval practice, we fit conditional probabilities: first test accuracy (see Figure 1A) and second test accuracy broken down separately for words pairs for which the first test was correct or incorrect, with this breakdown examined for second tests in the same direction (see Figure 1B) versus reverse direction (see Figure 1C). To increase reliability, we collapsed the data across a manipulation of how many times each word pair was initially studied and collapsed across screen position (e.g., the condition with both tests in the forward direction was combined with the condition with both tests in the backward direction). Another limitation of the original model was use of least-squares goodness of fit, which neglects small differences near the extremes of 0 and 1. Instead, we used G^2 as determined from maximum likelihood fits, which better respects the bounded accuracy scale (Riefer & Batchelder, 1988). After running into local minima problems with several different optimization routines (Nelder & Mead, 1965; Shi & Eberhart, 1998), we used brute force grid search in

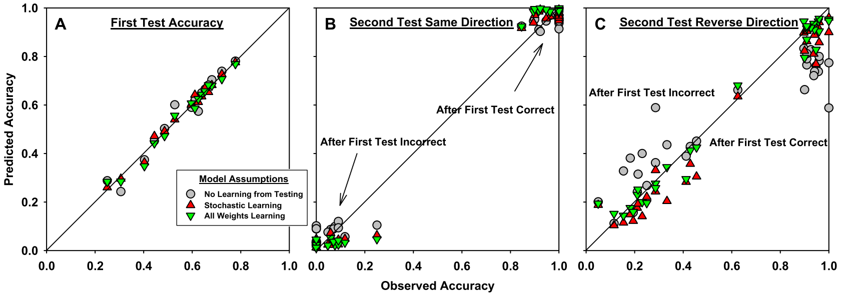

Figure 1: Comparison between observed and predicted accuracy when fitting three models with different assumptions about retrieval practice (no learning, stochastic learning, and all-weights learning) to the results from Kahana (2002). All models captured first test accuracy (A) and second test accuracy in the same direction as the first test (B), but the model without retrieval practice learning was unable to capture second test accuracy in the reverse direction (C). Each symbol shows the results for an individual participant (average of 36 data points for observed and 3000 for model). See https://github.com/asneha213/Paired-associate-learning for model code.

steps of .1 across the full 0 to 1 range of each parameter value, followed by a simplex search using the best results from the grid search.

We fit three models to the conditional accuracy values. The no learning from testing model contained three free parameters for each participant, capturing variation across word pairs (μ, σ, ρ), and two different retrieval practice models added the parameter μ_t to capture additional learning from test trials. In the stochastic retrieval practice learning model, upon successful recall, all weights, regardless of direction, were independently updated with probability μ_t. Of note, this learning is comparable to the stochastic learning that occurred between nodes of the same word during initial study of the word pairs, which should be on average symmetric. We also examined an all-weights retrieval practice learning model that enforced fully symmetric retrieval practice learning by updating all weights of the word pair. Thus, neither retrieval practice learning model introduced asymmetries, but they might accentuate existing asymmetries created during initial study of a word pair. An initial exploration of the all-weights model used μ_t in an all-or-none update, such that with probability μ_t, all weights were updated with a learning rate of 1.0 upon a correct first test. In replication of the original study, this model did not fit better than the model without retrieval practice. We determined that when all weights were updated with a learning rate of 1.0, this created a super-strong attractor that produced catastrophic interference for other word pairs (McCloskey & Cohen, 1989). In other words, retrieval practice did more harm than good with this magnitude of learning. To solve this problem, the all-weights retrieval practice

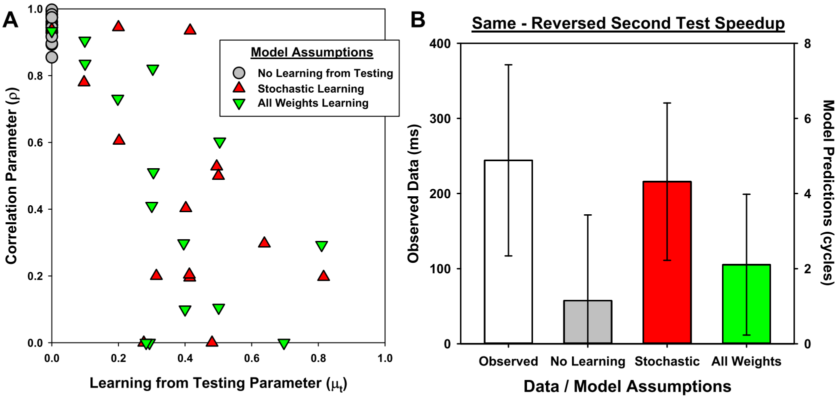

Figure 2: Model results suggesting asymmetric learning. Best-fitting correlation values are shown for each of the three models (A), with the learning from the testing parameter set to 0 for the no learning from testing model. Participants who were better fit with learning from testing ($\mu_t > 0$) were better fit with small correlation values ($\rho \ll 1$). A new analysis of the recall latencies (B) revealed that the speedup for a second test compared to a first test was greater when the second test was in the same direction as compared to the reverse direction ($p < .05$). Using model parameters that best fit the accuracy data, the model with stochastic retrieval practice learning predicted this latency effect ($p < .05$).

learning model was modified such that all weights were deterministically updated upon correct recall and μ_t was instead the learning rate ($\mu_t < 1$).

In replication of the original modeling study, the average best-fitting ρ value was .94 for the model without retrieval practice. However, goodness of fit for both retrieval practice models was nearly half that of the no learning from testing model (average goodness of fit: $G^2_{\text{no-learning}} = 9.1$; $G^2_{\text{stochastic}} = 4.6$; $G^2_{\text{all-weights}} = 4.9$). This occurred because the model without retrieval practice did not capture the reversed second test results (see Figure 1C; see also appendix B). The better-fitting retrieval practice models produced much lower correlation values (average: $\rho_{\text{stochastic}} = .45$; $\rho_{\text{all-weights}} = .44$).

The important theoretical question is whether newly learned word pairs take on stronger associations in one direction compared to the other. If so, the best-fitting correlation values should be much lower than 1.0. Regarding the question of whether asymmetric learning occurs in some situations, the average correlation values reported above ignore individual differences (see Figure 2A), with best-fit values covering the full range from 0 to 1. As seen in Figure 2A, there was a negative relationship between learning from testing and the best-fitting correlation value: participants who did not require retrieval practice were best fit with correlations near 1, whereas participants who exhibited retrieval practice effects were best fit by small correlations or even correlations equal to zero (the fitting routine did not allow for negative correlations). Considering that cued recall is a directional test, it is unclear if retrieval practice caused this asymmetry or if retrieval practice accentuated existing asymmetries, but in either case, these results provide clear evidence of correlations less than 1, supporting the existence of asymmetric learning.

Recall latency can uncover hidden benefits of retrieval practice and may serve to highlight asymmetric aspects of learning. For instance, after a first cued recall test without correct answer feedback, such as was the case in the Kahana (2002) experiment, recall latency decreases on a second cued recall test of the same word pairs even if accuracy is unchanged (Hopper & Huber, 2018, 2019). Therefore, we reanalyzed the original data for recall latency (Figure 2B), finding that second tests were correctly recalled more quickly after a correct first test, but only when the second test was in the same direction. Only the stochastic retrieval practice model predicted this effect (see also appendix C).

Other studies have reported that the benefits of cued recall practice are highly specific and directional (Hopper & Huber, 2018; Pan, Wong, Potter, Mejia, & Rickard, 2016), particularly when examining recall latency (Popov, Zhang, Koch, Calloway, & Coutanche, 2019). For instance, Popov et al. (2019) alternated test direction for semantically related word pairs (e.g., study: STUDENT-DORMITORY, test 1: STUDENT-?, test 2: ?-DORMITORY. test 3: STUDENT- ?, test 4: ? - DORMITORY). Remarkably, latency did not decrease at all on consecutive tests in the opposite direction and even tended to increase slightly, and yet latency decreased substantially on every other test (i.e., between tests in the same direction). For related word pairs, asymmetry might be learned over a lifetime of using them in a specific, directed manner to represent different concepts (e.g., "student dormitory" is a kind of building whereas "dormitory student" is a kind of person). Perhaps the modest asymmetry with unrelated word pairs marks the inception of this asymmetry, with asymmetry growing with repeated access of the word pair in a directed manner.

In conclusion, a reanalysis of the original modeling study and a new analysis of the original recall latency results suggest that retrieval practice is crucial for a full explanation of the data and that learning of unrelated word pairs involves a relatively high degree of asymmetry.

Appendix A: Additional Simulation Details

Weights were initially set to 0 before initial study of word pairs. Equation A.1 shows the stochastic Hopfield update equation for a pair of nodes, i and j, with the activation state values S_i and S_j set to appropriate $+1$ or -1 values as dictated by the word representations for a particular word pair presented for study. This equation was applied separately to the directed weight from node j to i, and vice versa from node i to node j, with the probability ϕ determining whether weight updating occurred in each direction. Self-weights (e.g., from node i to node i) were included:

$$\Delta W_{ij} = \begin{cases} \xi S_i S_j & \text{with probability } \phi \\ 0 & \text{with probability } 1 - \phi \end{cases}. \tag{A.1}$$

The learning rate ξ was set to 1.0 for all models except for the all-weights retrieval practice learning model, where it was set to μ_t for the updating that occurred after a successful recall. The stochastic update probability ϕ was set to μ for weights connecting nodes within the same word in response to initial study of a word pair and set to μ_t for learning after correct recall in the stochastic retrieval practice model. For each simulated word pair, a sample was taken from a bivariate normal distribution, with forward/backward means set to μ, forward/backward standard deviations set to σ, and correlation ρ to determine the separate ϕ update probabilities in each direction. If either of the two forward/backward ϕ values sampled from the bivariate normal was outside the 0 to 1 range, a new sample was taken, with this repeated until within-range values were obtained. These two forward/backward ϕ probabilities were then used to determine update probabilities between nodes belonging to different words of the word pair. In the case of the all-weights retrieval practice learning model, ϕ was set to 1.0 (deterministic updating) for learning after a correct recall.

To implement a cued recall trial, the 70 activation state values corresponding to cue word were fixed to $+1/-1$ values dictated by the word. The remaining 70 activation state values corresponding to the target word were initialized to random $+1/-1$ values. At each time step t, one of the 70 target nodes, i, was randomly selected (sampling with replacement), and the activation of that node was updated according to equation A.2, which sums the weighted inputs to node i from all 140 nodes, setting the activation to $+1$ or -1 for time step $t+1$ depending on the mathematical sign of the summed input. This updating of one node at a time occurred until either the correlation between the target nodes and the target exceeded .99, at which time the target was deemed to have been recalled, or 800 time steps elapsed, at which point the recall attempt was deemed a failure.

$$S_i(t+1) = sign\left(\sum_j W_{ij} S_j(t)\right). \tag{A.2}$$

Appendix B: Why Retrieval Practice Is Necessary

The model without retrieval practice had three free parameters, but there were four key aspects of the data: first test accuracy, extreme dependence for second test accuracy in the same direction, and two different conditional accuracy values for second test accuracy in the reverse direction. All models captured individual differences in first test accuracy (see Figure 1A) by setting the μ parameter to different values for different individuals. In contrast, the extreme dependency of second test accuracy for a second test in the same direction (see Figure 1B) was captured by different models in different ways. The model without retrieval practice captured this extreme dependency by setting the between word-pair variability parameter (σ) to a high value to produce a strong word-pair selection effect (a word pair producing an accurate first test was likely a well-studied word pair and thus likely to produce an accurate second test). In contrast, the retrieval practice learning models explained this extreme dependency as resulting from the retrieval practice parameter μ_t (after a correct first test, retrieval practice learning ensured a correct second test).

Whether each model could capture the two different conditional accuracy values for a reverse-direction second test (see Figure 1C: after a correct first test versus after an incorrect first test) was a matter of how many parameters remained. For the retrieval practice models, both the correlation parameter (ρ) and the word pair variability parameter were exploited to capture these two different values, but for the model without retrieval practice, the correlation parameter was the only available parameter. As a result, the model without retrieval practice captured one or the other of these values but not both. Thus, the high correlation values from the model without retrieval practice were the spurious by-product of a model that was qualitatively unable to capture the data. More specifically, because the word-pair variability parameter needed to be high for this model, the correlation parameter needed to be high as well to maintain an appropriate level of conditional variance (conditional variance of a bivariate normal distribution is equal to the variance times one minus correlation squared).

Appendix C: Descriptions of Latency Effects

The recall latency analysis was run in a pair-wise manner by first selecting all word pairs for which both test trials were correct (both data and model) and then taking the difference in milliseconds (data) or time steps (model) between a correct first test and a correct second test. For each participant, the average of these difference scores for a reverse-direction second test was subtracted from the average difference score for a same-direction second test. Thus, the results shown in Figure 2B are a difference of differences (an interaction effect), and the significance values reported in the figure caption reflect paired samples t-tests against value 0 for this difference of differences.

For the model without retrieval practice, the first test does not affect the second test, and so the order of the tests is irrelevant (a systematically faster second test is not possible if the order of tests is irrelevant). In contrast, both retrieval practice models predicted a speedup on the second test, although this speedup was not equivalent between a same- versus reverse-direction second test. Because retrieval practice learning occurs only after a correct first test (i.e., it only applies to target words that were already recallable), it does not increase accuracy on a second test in the same direction, but it does decrease recall latency. In contrast, retrieval practice might increase accuracy on a reverse-direction second test, particularly if the correlation is low. However, an increase in accuracy comes at the expense of the latency effect because words that are converted from unrecallable to just barely recallable are recalled slowly. The extent to which this occurred in the reverse direction was slightly different for the two kinds of retrieval practice learning (i.e., strong updating for a few weights versus weak updating for all weights). Nevertheless, this was a subtle effect, and as seen in Figure 2B, the recall latency difference between the stochastic model and the all-weights model is not reliable.

References

Abbott, E. E. (1909). On the analysis of the factor of recall in the learning process. *Psychological Monographs*, *11*(1), 159–177. 10.1037/h0093018

Bi, G. Q., & Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, *18*(24), 10464–10472. 10.1523/JNEUROSCI.18-24-10464.1998, PubMed: 9852584

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual-cortex. *Journal of Neuroscience*, *2*(1), 32–48. 10.1523/JNEUROSCI.02-01-00032.1982, PubMed: 7054394

Cox, G. E., & Criss, A. H. (2020). Similarity leads to correlated processing: A dynamic model of encoding and recognition of episodic associations. *Psychological Review*, *127*(5), 792–828. 10.1037/rev0000195, PubMed: 32191075

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the National Academy of Sciences of the United States of America–Biological Sciences*, *79*(8), 2554–2558.

Hopper, W. J., & Huber, D. E. (2018). Learning to recall: Examining recall latencies to test an intra-item learning theory of testing effects. *Journal of Memory and Language*, *102*, 1–15. 10.1016/j.jml.2018.04.005

Hopper, W. J., & Huber, D. E. (2019). Testing the primary and convergent retrieval model of recall: Recall practice produces faster recall success but also faster recall failure. *Memory and Cognition*, *47*(4), 816–841. 10.3758/s13421-019-00903-x, PubMed: 30737729

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory and Cognition*, *30*(6), 823–840. 10.3758/BF03195769, PubMed: 12450087

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, *24*, 109–165. 10.1016/S0079-7421(08)60536-8

Nelder, J. A., & Mead, R. (1965). A Simplex method for function minimization. *Computer Journal*, *7*, 308–313. 10.1093/comjnl/7.4.308

Pan, S. C., Wong, C. M., Potter, Z. E., Mejia, J., & Rickard, T. C. (2016). Does test-enhanced learning transfer for triple associates? *Memory and Cognition*, *44*(1), 24–36. 10.3758/s13421-015-0547-x, PubMed: 26324093

Pezdek, K., Roman, Z., & Sobolik, K. G. (1986). Spatial memory for objects and words. *Journal of Experimental Psychology-Learning Memory and Cognition*, *12*(4), 530–537. 10.1037/0278-7393.12.4.530

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. 10.1037/a0014420, PubMed: 19159151

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127*(1), 1–46. 10.1037/rev0000161, PubMed: 31524424

Popov, V., Zhang, Q., Koch, G. E., Calloway, R. C., & Coutanche, M. N. (2019). Semantic knowledge influences whether novel episodic associations are represented symmetrically or asymmetrically. *Memory and Cognition*, *47*(8), 1567–1581. 10.3758/s13421-019-00950-4, PubMed: 31215011

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive-processes. *Psychological Review*, *95*(3), 318–339. 10.1037/0033-295X.95.3.318

Rizzuto, D. S., & Kahana, M. J. (2001). An autoassociative neural network model of paired-associate learning. *Neural Computation*, *13*(9), 2075–2092. 10.1162/089976601750399317, PubMed: 11516358

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. 10.1016/j.tics.2010.09.003, PubMed: 20951630

Shi, Y. H., & Eberhart, R. (1998). A modified particle swarm optimizer. In *Proceedings of the 1998 IEEE International Conference on Evolutionary Computation* (pp. 69–73). Piscataway, NJ: IEEE. 10.1109/Icec.1998.699146

Received April 7, 2021; accepted June 14, 2021.