LETTER Communicated by Stephen José Hanson

Free-Lunch Learning: Modeling Spontaneous Recovery of Memory

J. V. Stone
j.v.stone@sheffield.ac.uk
Psychology Department, Sheffield University, Sheffield S10 2TP, England

P. E. Jupp
pej@st-andrews.ac.uk
School of Mathematics and Statistics, St. Andrews University, St. Andrews KY16 9SS, Scotland

After a language has been learned and then forgotten, relearning some words appears to facilitate spontaneous recovery of other words. More generally, relearning partially forgotten associations induces recovery of other associations in humans, an effect we call *free-lunch learning* (FLL). Using neural network models, we prove that FLL is a necessary consequence of storing associations as distributed representations. Specifically, we prove that (1) FLL becomes increasingly likely as the number of synapses (connection weights) increases, suggesting that FLL contributes to memory in neurophysiological systems, and (2) the magnitude of FLL is greatest if inactive synapses are removed, suggesting a computational role for synaptic pruning in physiological systems. We also demonstrate that FLL is different from generalization effects conventionally associated with neural network models. As FLL is a generic property of distributed representations, it may constitute an important factor in human memory.

1 Introduction

A popular aphorism states that "there's no such thing as a free lunch." However, in the context of learning theory, we propose that there is. In previous work, free-lunch learning (FLL) has been demonstrated using a task in which participants learned the positions of letters on a nonstandard computer keyboard (Stone, Hunkin, & Hornby, 2001). After a period of forgetting, participants relearned a proportion of these letter positions. Crucially, it was found that this relearning induced recovery of the nonrelearned letter positions. Preliminary results suggest that FLL also occurs using face stimuli.

If the brain stores information as distributed representations, then each neuron contributes to the storage of many associations, so that relearning

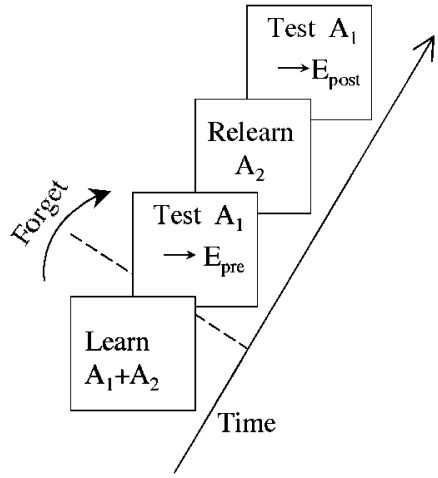


Figure 1: Free-lunch learning protocol. Two subsets of associations A_1 and A_2 are learned. After partial forgetting (see text), performance error E_{pre} on subset A_1 is measured. Subset A_2 is then relearned to preforgetting levels of performance, and performance error E_{post} on subset A_1 is remeasured. If $E_{post} < E_{pre}$ then FLL has occurred, and the amount of FLL is $\delta = E_{pre} - E_{post}$.

some old and partially forgotten associations affects the integrity of other old associations. Using neural network models, we show that relearning some associations does not disrupt other stored associations but actually restores them.

In essence, recovery occurs in neural network models because each association is distributed among all connection weights (synapses) between units (model neurons). After partial forgetting, relearning some of the associations forces all of the weights closer to preforgetting values, resulting in improved performance even on nonrelearned associations.

1.1 The Geometry of Free-Lunch Learning. The protocol used to examine FLL here is as follows (see Figure 1). First, learn a set of $n_1 + n_2$ associations $A = A_1 \cup A_2$ consisting of two intermixed subsets A_1 and A_2 of n_1 and n_2 associations, respectively. After all learned associations A have been partially forgotten, measure performance on subset A_1. Finally, relearn only subset A_2, and then remeasure performance on subset A_1. FLL occurs if relearning subset A_2 improves performance on A_1. Unless stated otherwise, we assume that for a network with n connection weights, $n \geq n_1 + n_2$.

For the present, we assume that the network has one output unit and two input units, which implies $n = 2$ connection weights and that A_1 and A_2 each consist of $n_1 = n_2 = 1$ association, as in Figure 2. Input units are

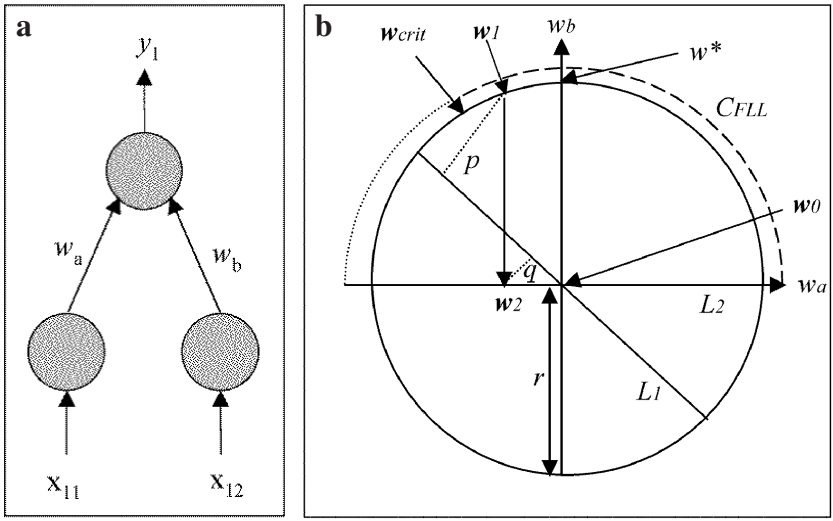


Figure 2: Geometry of free-lunch learning. (a) A network with two input units and one output unit, with connection weights w_a and w_b, defines a weight vector $\mathbf{w} = (w_a, w_b)$. The network learns two associations A_1 and A_2, where (for example) A_1 is the mapping from input vector $\mathbf{x}_1 = (x_{11}, x_{12})$ to desired output value d_1; learning consists of adjusting $\mathbf{w}$ until the network output $y_1 = \mathbf{w} \cdot \mathbf{x}_1$ equals d_1. (b) Each association A_1 and A_2 defines a constraint line L_1 and L_2, respectively. The intersection of L_1 and L_2 defines a point $\mathbf{w}_0$ that satisfies both constraints, so that zero error on A_1 and A_2 is obtained if $\mathbf{w} = \mathbf{w}_0$. After partial forgetting, $\mathbf{w}$ is a randomly chosen point $\mathbf{w}_1$ on the circle C with radius r, and performance error E_{pre} on A_1 is the squared distance p^2. After relearning A_2, the weight vector $\mathbf{w}_2$ is in L_2, and performance error E_{post} on A_1 is q^2. FLL occurs if $\delta = E_{pre} - E_{post} > 0$, or equivalently if $Q = p^2 - q^2 > 0$. Relearning A_2 has one of three possible effects, depending on the position of $\mathbf{w}_1$ on C: (1) if $\mathbf{w}_1$ is under the larger (dashed) arc C_{FLL} as shown here, then $p^2 > q^2$ ($\delta > 0$) and therefore FLL is observed; (2) if $\mathbf{w}_1$ is under the smaller (dotted) arc, then $p^2 < q^2$ ($\delta < 0$), and therefore negative FLL is observed; and (3) if $\mathbf{w}_1$ is at the critical point $\mathbf{w}_{crit}$, then $p^2 = q^2$ ($\delta = 0$). Given that $\mathbf{w}_1$ is a randomly chosen point on C and that the length of C_{FLL} is S_{FLL}, the probability of FLL is $P(\delta > 0) = S_{FLL}/\pi r$ (i.e., the proportion of C_{FLL} under the upper semicircle of C).

connected to the output unit via weights w_a and w_b, which define a weight vector $\mathbf{w} = (w_a, w_b)$. Associations A_1 and A_2 consist of different mappings from the input vectors $\mathbf{x}_1 = (x_{11}, x_{12})$ and $\mathbf{x}_2 = (x_{21}, x_{22})$ to desired output values d_1 and d_2, respectively. If a network is presented with input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, then its output values are $y_1 = \mathbf{w} \cdot \mathbf{x}_1 = w_a x_{11} + w_b x_{12}$ and $y_2 = \mathbf{w} \cdot \mathbf{x}_2 = w_a x_{21} + w_b x_{22}$, respectively. Network performance error for $k = 2$ associations is defined as $E(\mathbf{w}, A) = \sum_{i=1}^{k} (d_i - y_i)^2$.

The weight vector $\mathbf{w}$ defines a point in the (w_a, w_b)-plane. For an input vector $\mathbf{x}_1$, there are many different combinations of weight values w_a and w_b that give the desired output d_1. These combinations lie on a straight line L_1, because the network output is a linear weighted sum of input values. A corresponding constraint line L_2 exists for A_2. The intersection of L_1 and L_2 therefore defines the only point $\mathbf{w}_0$ that satisfies both constraints, so that zero error on A_1 and A_2 is obtained if and only if $\mathbf{w} = \mathbf{w}_0$. Without loss of generality, we define the origin $\mathbf{w}_0$ to be the intersection of L_1 and L_2.

We now consider the geometric effect of partial forgetting of both associations, followed by relearning A_2. This geometric account applies to a network with two weights (see Figure 2) and depends on the following observation: if the length of the input vector $\|\mathbf{x}_1\| = 1$, then the performance error $E(\mathbf{w}, A_1) = (d_1 - y_1)^2$ of a network with weight vector $\mathbf{w}$ when tested on association A_1 is equal to the squared distance between $\mathbf{w}$ and the constraint line L_1 (see appendix C). For example, if $\mathbf{w}$ is in L_1, then $E(\mathbf{w}, A_1) = 0$, but as the distance between $\mathbf{w}$ and L_1 increases, so $E(\mathbf{w}, A_1)$ must increase. For the purposes of this geometric account, we assume that $\|\mathbf{x}_1\| = \|\mathbf{x}_2\| = 1$.

Partial forgetting is induced by adding isotropic noise $\mathbf{v}$ to the weight vector $\mathbf{w} = \mathbf{w}_0$. This effectively moves $\mathbf{w}$ to a randomly chosen point $\mathbf{w}_1 = \mathbf{w}_0 + \mathbf{v}$ on the circle C of radius $r = \|\mathbf{v}\|$, where r represents the amount of forgetting. For a network with $\mathbf{w} = \mathbf{w}_1$, learning A_2 moves $\mathbf{w}$ to the nearest point $\mathbf{w}_2$ on L_2 (see appendix B), so that $\mathbf{w}_2$ is the orthogonal projection of $\mathbf{w}_1$ on L_2. Before relearning A_2, the performance error E_{pre} on A_1 is the squared distance p^2 between $\mathbf{w}_1$ and its orthogonal projection on L_1 (see appendix C). After relearning A_2, the performance error E_{post} is the squared distance q^2 between $\mathbf{w}_2$ and its orthogonal projection on L_1. The amount of FLL is $\delta = E_{pre} - E_{post}$ and, for a network with two weights, is equal to $Q = p^2 - q^2$. The probability $P(\delta > 0)$ of FLL given L_1 and L_2 is equal to the proportion of points on C for which $\delta > 0$ (or, equivalently, for which $Q > 0$). For example, averaging over all subsets A_1 and A_2, there is the probability $P(\delta > 0) = 0.68$ that relearning A_2 induces FLL of A_1 (see Figure 5), a probability that increases with the number of weights (see theorem 3).

If we drop the assumption that a network has only two input units, then we can consider subsets A_1 and A_2 with $n_1 > 1$ and $n_2 > 1$ associations. If the number of connection weights $n \geq \max(n_1, n_2)$, then A_1 and A_2 define an $(n - n_1)$-dimensional subspace L_1 and an $(n - n_2)$-dimensional subspace L_2, respectively. The intersection of L_1 and L_2 corresponds to weight vectors that generate zero error on $A = A_1 \cup A_2$.

Finally, we can drop the assumption that a network has only one output unit, because the connections to each output unit can be considered as a distinct network, in which case our results can be applied to the network associated with each output unit.

2 Methods

Given a network with n input units and one output unit, the set A of associations consisted of k input vectors $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ and k corresponding desired scalar output values $(d_1, \ldots, d_k)$. Each input vector comprises n elements $\mathbf{x} = (x_1, \ldots, x_n)$. The values of x_i and d_i were chosen from a gaussian distribution with unit variance (i.e., $\sigma_x^2 = \sigma_d^2 = 1$). A network's output y_i is a weighted sum of input values $y_i = \mathbf{w} \cdot \mathbf{x}_i = \sum_{j=1}^{k} w_j x_{ij}$, where x_{ij} is the jth value of the ith input vector $\mathbf{x}_i$, and each weight w_i is one input-output connection.

Given that the network error for a given set of k associations is $E(\mathbf{w}, A) = \sum_{i=1}^{k}(d_i - y_i)^2$, the derivative $\nabla E_{(\mathbf{w})} = 2\sum_{i=1}^{k}(d_i - y_i)\mathbf{x}_i$ of E with respect to $\mathbf{w}$ yields the delta learning rule $\mathbf{w}_{new} = \mathbf{w}_{old} - \eta \nabla E_{(\mathbf{w}_{old})}$, where η is the learning rate, which is adjusted according to the number of weights. A learning trial consists of presenting the k input vectors to the network and then updating the weights using the delta rule. Learning was stopped when $\|\nabla E_{(\mathbf{w})}\| < k0.001$, where $\|\nabla E_{(\mathbf{w})}\|$ is the magnitude of the gradient.

Initial learning of the $k = n$ associations in $A = A_1 \cup A_2$ was performed by solving a set of n simultaneous equations using a standard method, after which perfect performance on all n associations was obtained. Partial forgetting was induced by adding an isotropic noise vector $\mathbf{v}$ with $r = \|\mathbf{v}\| = 1$. Relearning the $n_2 = n/2$ associations in A_2 was implemented with $k = n_2$ using the delta rule.

3 Results

Our four main theorems are summarized here, and proofs are provided in the appendixes. These theorems apply to a network with n weights that learns $n_1 + n_2$ associations $A = A_1 \cup A_2$ and, after partial forgetting, relearns the n_2 associations in A_2.

Theorem 1. *The probability $P(\delta > 0)$ of FLL is greater than 0.5.*

Theorem 2. *The expected amount of FLL per association in A_1 is*

$$E[\delta/n_1] = \frac{n_2}{n^2} E[\|\boldsymbol{x}\|^2] E[\|\boldsymbol{v}\|^2]. \tag{3.1}$$

For given values of $\mathrm{E}[\|\mathbf{x}\|^2]$ and $\mathrm{E}\left[\|\mathbf{v}\|^2\right]$, the value of n_2, which maximizes $\mathrm{E}[\delta/n_1]$ (subject to $n_1 + n_2 \leq n$), is $n_2 = n - n_1$.

If each input vector $\mathbf{x} = (x_1, \ldots, x_n)$ is chosen from an isotropic (e.g., isotropic gaussian) distribution and the variance of x_i is σ_x^2, then $\mathrm{E}\left[\|\mathbf{x}\|^2\right] = n\sigma_x^2$. If σ_x^2 is the same for all n, then the state of a neuron (with a typical

sigmoidal transfer function) would be in a constantly saturated state as the number of synapses increases. One way to prevent this saturation is to assume that the efficacy of synapses on a given neuron decreases as the number of synapses increases. If forgetting is caused primarily by learning spurious inputs, then the delta learning rule used here implies that the "amount of forgetting" $\|\mathbf{v}\|$ is approximately independent of n. We therefore assume that $\|\mathbf{v}\|$ and σ_x^2 are constant, and for convenience, we set $\|\mathbf{v}\| = 1$ and $\sigma_x^2 = 1$. Substituting these values into equation 3.1 yields

$$\mathrm{E}[\delta/n_1] = \frac{n_2}{n}. \tag{3.2}$$

Using these assumptions, simulations of networks with $n = 2$ and $n = 100$ weights agree with equation 3.2, as shown in Figure 3.

The role of pruning can be demonstrated as follows. Consider a network with 100 input units and one output unit with $n = 100$ weights. If $n_2 = 90$ associations are relearned out of an original set of $n_1 + n_2 = 100$ associations, then $\mathrm{E}[\delta/n_1] = n_2/n = 0.90$. However, if $n = 1000$, then $\mathrm{E}[\delta/n_1] = 0.09$. In general, as the number $n - (n_1 + n_2)$ of unpruned redundant weights increases, so $\mathrm{E}[\delta/n_1]$ decreases. Therefore, $\mathrm{E}[\delta/n_1]$ is maximized if $n_1 + n_2 = n$. If $n_1 + n_2 < n$, then the expected amount of FLL is not maximal and can therefore be increased by pruning redundant weights until $n = n_1 + n_2$ (see Figure 4).

Note that for a particular network, performance error E_{post} on A_1 after learning A_2 can be zero. For example, if $\mathbf{w} = \mathbf{w}^*$ in Figure 2, then $p = q = 0$, which implies that $\delta/n_1 = E_{post} = q^2 = 0$.

Theorem 3. *The probability $P(\delta > 0)$ of FLL of A_1 satisfies*

$$P(\delta > 0) > 1 - \frac{a_0(n, n_1, n_2) + a_1(n, n_2)\, var\left(\|\boldsymbol{x}\|^2\right)/E\left[\|\boldsymbol{x}\|^2\right]^2}{n_1 n_2 (n+2)^2}, \tag{3.3}$$

where

$$a_0(n, n_1, n_2) = 2\left\{n_1(n+2)(n-n_2) + n(n-n_2) + n(n+2)(n-1)\right\} \tag{3.4}$$

$$a_1(n, n_2) = n^2(2n + n_2 + 6). \tag{3.5}$$

Theorem 3 implies that if the numbers (n_1 and n_2) of associations in A_1 and A_2 are fixed nonzero proportions of the number n of connection weights (n_1/n and n_2/n, respectively) and $\mathrm{var}\left(\|\mathbf{x}\|^2\right)/n\mathrm{E}\left[\|\mathbf{x}\|^2\right]^2 \to 0$ as $n \to \infty$, then $P(\delta > 0) \to 1$ as $n \to \infty$; and the probability that each of the n_1 associations in A_1 exhibits FLL is $P(\delta/n_1 > 0) = P(\delta > 0)$ because $\delta > 0$ iff $\delta/n_1 > 0$.

For example, if we assume that each input vector $\mathbf{x} = (x_1, \ldots, x_n)$ is chosen from an isotropic (e.g., isotropic gaussian) distribution and the

variance of x_i is σ_x^2, then $\mathrm{var}\left(\|\mathbf{x}\|^2\right)/\mathrm{E}\left[\|\mathbf{x}\|^2\right]^2 = 2/n$. This ensures that $\mathrm{var}\left(\|\mathbf{x}\|^2\right)/n\mathrm{E}\left[\|\mathbf{x}\|^2\right]^2 \to 0$ as $n \to \infty$, and therefore that $P(\delta > 0) \to 1$ as $n \to \infty$.

Using this assumption, an approximation of the right-hand side of equation 3.3 yields

$$P(\delta > 0) > 1 - \frac{2(1 + \alpha_1 - \alpha_1\alpha_2)}{n\alpha_1\alpha_2} - \frac{2(2 + \alpha_2 + 6/n)}{\alpha_1\alpha_2(n+2)^2}, \tag{3.6}$$

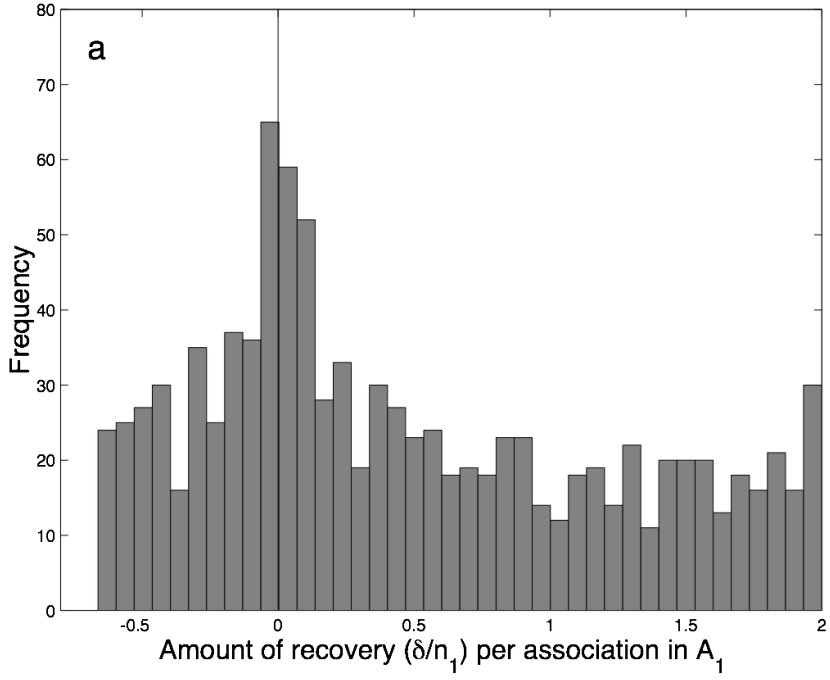
where $\alpha_1 = n_1/n$ and $\alpha_2 = n_2/n$. In this form, it is easy to see that $P(\delta > 0) \to 1$ as $n \to \infty$.

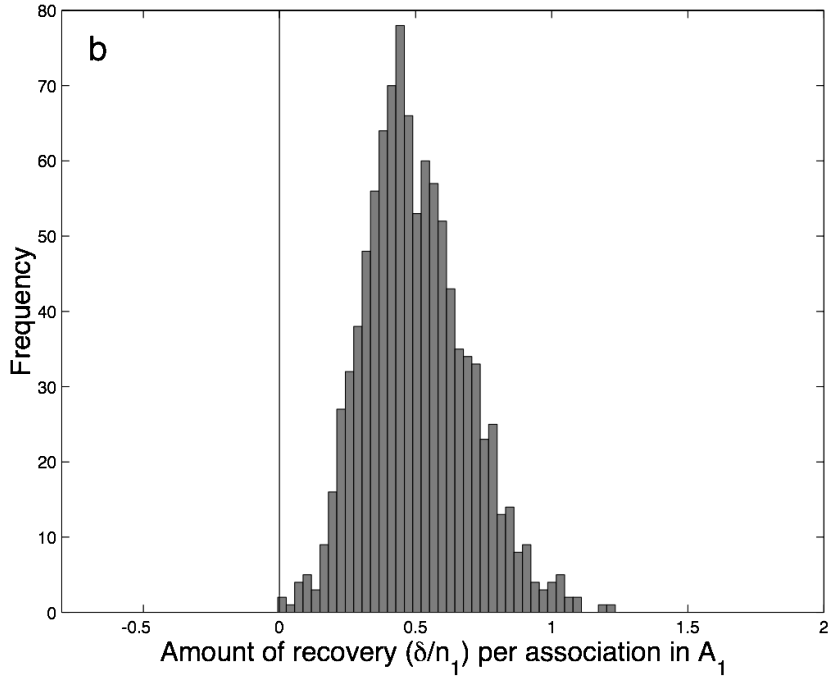
We briefly consider the case $n_1 \geq n$ and $n_2 \geq n$, so that each of L_1 and L_2 is a single point. If the distance D between these points is much less than $\|\mathbf{v}\|$, then simple geometry shows that performance error E_{pre} on A_1 is large and that relearning A_2 reduces this error for any $\mathbf{v}$ (i.e., with probability 1) with $E_{post} \propto D^2$, even in the absence of initial learning of A_1 and A_2 (see equation A.18 in appendix A). A similar conclusion is implicit in Atkins and Murre (1998).

Theorem 4. *If, instead of relearning A_2, the network learns a new subset A_3 (drawn from the same distribution as A_2), then the expected amount of FLL is less than the expected amount of FLL after relearning subset A_2.*

Learning A_3 is analogous to the control condition used with human participants (Stone et al., 2001), and the finding that the amount of recovery after learning A_3 is less than the amount of recovery after relearning A_2 is predicted by theorem 4.

Figure 3: Distribution of free-lunch learning. (a) Histogram of amount of FLL δ/n_1 per association, based on 1000 runs, for a network with $n = 2$ weights (see section 2). After learning two association subsets ($\eta = 0.1$), A_1 and A_2, containing $n_1 = 1$ and $n_2 = 1$ associations (respectively), the network has a weight vector $\mathbf{w}_0$. Forgetting is then induced by adding a noise vector $\mathbf{v}$ with $\|\mathbf{v}\|^2 = 1$ to $\mathbf{w}_0$. One association A_2 is then relearned, and the change in performance on A_1 is measured as δ/n_1 (see Figure 2). Negative values indicate that performance on A_1 decreases after relearning A_2. (b) Histogram of amount of FLL δ/n_1 per association for a network with $n = 100$ weights and $\eta = 0.005$, with A_1 and A_2 each consisting of $n_1 = n_2 = 50$ associations, using the same protocol as in (a). In both (a) and (b), the mean value of δ/n_1 is about 0.5, as predicted by equation 3.2. As the number of associations learned increases, the amount of FLL becomes more tightly clustered around $\delta/n_1 = 0.5$, as demonstrated in these two histograms, and the probability of FLL increases (also see Figure 5).


a
Frequency
Amount of recovery (δ/n1) per association in A1
0
10
20
30
40
50
60
70
80
-0.5
0
0.5
1
1.5
2


b
Frequency
Amount of recovery (δ/n1) per association in A1
0
10
20
30
40
50
60
70
80
-0.5
0
0.5
1
1.5
2

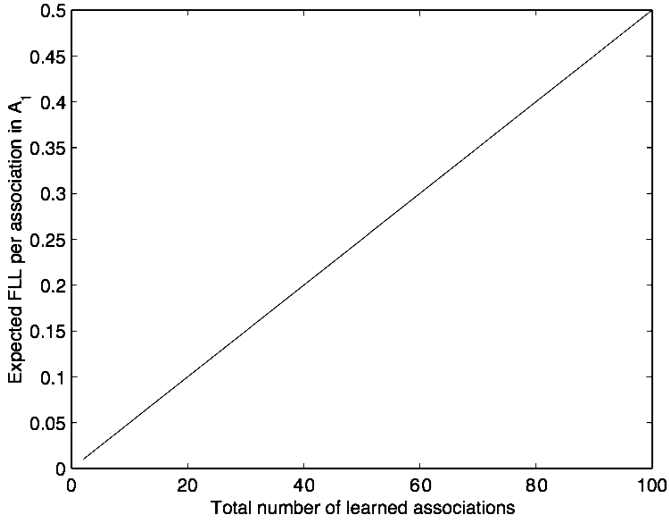


Figure 4: Effect of pruning on free-lunch learning. Graph of the expected amount of FLL per association $E[\delta/n_1]$ as a function of the total number $n_1 + n_2$ of learned associations in $A = A_1 \cup A_2$, as given in equation 3.2. In this example, the number of connection weights is fixed at $n = 100$, and the number of associations in $A = A_1 \cup A_2$ increases from $n_1 + n_2 = 2$ to $n_1 + n_2 = 100$. The number n_2 of relearned associations in A_2 is a constant proportion (0.5) of the associations in A. If $n_1 + n_2 \leq n$, then the network contains $n - (n_1 + n_2)$ unpruned redundant connections. Thus, pruning effectively increases as $n_1 + n_2$ increases because, as the number $n_1 + n_2$ of associations grows, so the number of unpruned redundant connections decreases. The expected amount of FLL per association $E[\delta/n_1]$ increases as the amount of pruning increases.

4 Discussion

Theorems 1 to 4 provide the first proof that relearning induces nontransient recovery, where postrecovery error is potentially zero. This contrasts with the usually small and transient recovery that occurs during the initial phase of relearning forgotten associations (Hinton & Plaut, 1987; Atkins & Murre, 1998), and during learning of new associations (Harvey & Stone, 1996). In particular, theorem 2 is predictive inasmuch as it suggests that the amount of FLL in humans should be (1) proportional to the amount of forgetting of $A = A_1 \cup A_2$ and (2) proportional to the proportion $n_2/(n_1 + n_2)$ of associations relearned after partial forgetting of A.

We have assumed that the number $n_1 + n_2$ of associations $A = A_1 \cup A_2$ encoded by a given neuron is not greater than the number n of input connections (synapses) to that neuron. Given that each neuron typically has

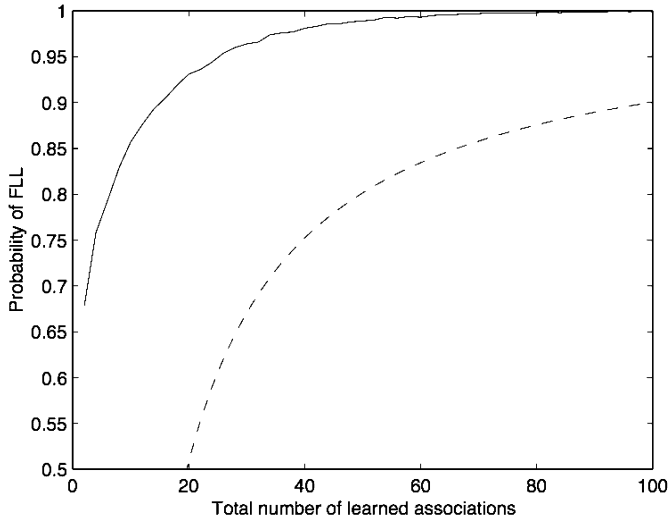


Figure 5: Probability of free-lunch learning. The probability $P(\delta > 0)$ of FLL of associations A_1 as a function of the total number $n_1 + n_2$ of learned associations $A = A_1 \cup A_2$ for networks with $n = n_1 + n_2$ weights. Each of the two subsets of associations A_1 and A_2 consists of $n_1 = n_2 = n/2$ associations. After learning and then partially forgetting A, performance on A_1 was measured. $P(\delta > 0)$ is the probability that performance on subset A_1 is better after subset A_2 has been relearned than it is before A_2 has been relearned. Solid line: Empirical estimate of $P(\delta > 0)$. Each data point is based on 10,000 runs, where each run uses input vectors chosen from an isotropic gaussian distribution (see section 2). Dashed line: Theoretical lower bound on the probability of FLL, as given by theorems 1 and 3, assuming that input vectors are chosen from an isotropic (e.g., isotropic gausssian) distribution.

many thousands of synapses (e.g., cerebellar Purkinje cells), it seems likely that this assumption is valid. However, the total amount of FLL is maximal if $n_1 = n_2 = n/2$, so that the full potential of FLL can be realized only if $n_1 + n_2 = n$. This optimum number of synapses can be achieved if inactive (i.e., redundant) synapses are pruned. Pruning may therefore contribute to FLL in physiological systems (Purves & Lichtman, 1980; Goldin, Segal, & Avignone, 2001).

We have also assumed that a delta rule is used to learn associations between inputs and desired outputs. This general type of supervised learning is thought to be implemented by the cerebellum and basal ganglia (Doya, 1999). Models of the cerebellum (Dean, Porrill, & Stone, 2002) use a delta rule to implement learning. Similarly, models of the basal ganglia (Nakahara, Itoh, Kawagoe, Takikawa, & Hikosaka, 2004) use a temporally discounted

form of delta rule, the temporal difference rule. This temporal difference rule has also been used to model learning in humans (Seymour et al., 2004), and (under mild conditions) is equivalent to the standard delta rule (Sutton, 1988). Indeed, from a purely computational perspective, it is difficult to conceive how these forms of associative learning could be implemented without some form of delta rule.

Our analysis is based on the assumption that the network model is linear. Of course, many nonlinear networks can be approximated by linear networks, but it is possible that the results derived here have limited applicability to certain classes of nonlinear networks.

Relation to Task Generalization. It is only natural to ask how FLL relates to tasks that a human might learn. One obvious but vital condition for FLL is that different associations must be encoded by a common set of neuronal connections. Aside from this condition, it might be thought that relearning A_2 improves performance on A_1 because A_1 and A_2 are somehow related (as in Hanson & Negishi, 2002; Dienes, Altmann, & Gao, 1999), so that learning A_2 generalizes to A_1. This form of *task generalization* can occur if A_1 and A_2 are related as follows. If the input-output pairs in A_1 and A_2 are sampled from a sufficiently smooth function f and $n_1 \gg n$ and $n_2 \gg n$, then A_1 and A_2 are statistically related, and therefore the weights induced by learning A_1 are similar to those induced by learning A_2. Consequently, the resultant network input-output functions g_1 and g_2 (respectively) both approximate the function f (i.e., $g_1 \approx g_2 \approx f$). In this case, learning A_2 yields good performance on A_1. In the context of FLL, if $A_1 \cup A_2$ is learned, forgotten, and then A_2 is relearned, performance on A_1 will also improve. However, the reason for this improvement is obvious and trivial: it is simply that A_1 and A_2 are statistically related and large enough (i.e., with $n_1 \gg n$ and $n_2 \gg n$) to induce similar network functions.

In contrast, the effect described in this letter does not depend on statistical similarity between A_1 and A_2. Crucial assumptions are that $n_1 + n_2 \leq n$, $n_1 < n$, and $n_2 < n$, so that learning the n_2 associations in A_2 in a network with n weights is underconstrained. This implies that the network function induced by learning A_1 has no particular relation to the network function induced by learning A_2, even if A_1 and A_2 are sampled from the same function f (provided A_1 and A_2 are disjoint sets). For example, if A_1 and A_2 each consists of one association sampled from a linear function f (i.e., a line), then learning A_2 in a linear network (as in Figure 2a) induces a linear network function g_1 (i.e., a line) that intersects with f but is otherwise unconstrained. Thus, learning A_2 does not necessarily yield good performance on A_1. The FLL effect reported here depends on relearning after forgetting. To cite an extreme example, if unicycling and learning French were encoded by a common set of neurons, then, after forgetting both, relearning unicycling could improve your French (although the mechanism involved here is unrelated to that described in Harvey & Stone, 1996). Thus, FLL contrasts

with the task generalization outlined above, where it is obvious that both A_1 and A_2 induce similar network functions.

Motivated by the demonstration that recovery occurs in humans (Stone et al., 2001; Coltheart & Byng, 1989; Weekes & Coltheart, 1996) (but not in all studies—Atkins, 2001), we have proven that FLL occurs in network models. The analysis presented here suggests that FLL is a necessary and generic consequence of storing information in distributed systems rather than a side effect peculiar to a particular class of artificial neural nets. Moreover, the generic nature of FLL suggests that it is largely independent of the type (i.e., artificial or physiological) of network used to learn associations.

FLL appears to be a fundamental property of distributed representations. Given the reliance of neuronal systems on distributed representations, FLL may be a ubiquitous feature of learning and memory. It is likely that any organism that did not take advantage of such a fundamental and ubiquitous effect would be at a severe selective disadvantage.

Appendix A: Analysis of Free-Lunch Learning

We proceed by deriving expressions for E_{pre}, E_{post}, and $\delta = E_{pre} - E_{post}$. We prove that if $n_1 + n_2 \leq n$, then the expected value of δ is positive. We then prove that if $n_1 + n_2 \leq n$, the probability $P(\delta > 0)$ of FLL is greater than 0.5, that its lower bound increases with n (if n_1/n and n_2/n are fixed), and that this bound approaches unity as n increases.

A.1 Definition of Performance Error. For an artificial neural network (ANN) with weight vector $\mathbf{w}$, we define the performance error for input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_c$ and desired outputs $d_1, \ldots, d_c$ to be

$$E(\mathbf{x}_1, \ldots, \mathbf{x}_c; \mathbf{w}, d_1, \ldots, d_c) = \sum_{i=1}^{c} (\mathbf{w} \cdot \mathbf{x}_i - d_i)^2 . \tag{A.1}$$

By putting $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_c)^T$, $\mathbf{d} = (d_1, \ldots, d_c)^T$ and

$$E(\mathbf{X}; \mathbf{w}, \mathbf{d}) = E(\mathbf{x}_1, \ldots, \mathbf{x}_c; \mathbf{w}, d_1, \ldots, d_c),$$

we can write equation A.1 succinctly as

$$E(\mathbf{X}; \mathbf{w}, \mathbf{d}) = \|\mathbf{X}\mathbf{w} - \mathbf{d}\|^2. \tag{A.2}$$

Given a $c \times n$ matrix $\mathbf{X}$ and a c-dimensional vector $\mathbf{d}$, let $L_{\mathbf{X},\mathbf{d}}$ be the affine subspace,

$$L_{\mathbf{X},\mathbf{d}} = \left\{ \mathbf{w} : \mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{d} \right\},$$

of $\mathbb{R}^n$. Since

i. $\mathrm{rk}\left(\mathbf{X}^T\mathbf{X}\right) \leq \mathrm{rk}\left(\mathbf{X}\right)$,
ii. $\mathbf{X}^T\mathbf{X}\mathbf{a} = \mathbf{0} \Rightarrow \mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{a} = 0 \Rightarrow \mathbf{X}\mathbf{a} = \mathbf{0}$,

it follows that $\mathrm{rk}\left(\mathbf{X}^T\mathbf{X}\right) = \mathrm{rk}\left(\mathbf{X}\right)$ (where rk denotes the rank of a matrix), and so

$$L_{\mathbf{X},\mathbf{d}} \text{ is nonempty.} \tag{A.3}$$

If $\mathbf{X}$ and $\mathbf{d}$ are consistent (i.e., there is a $\mathbf{w}$ such that $\mathbf{X}\mathbf{w} = \mathbf{d}$), then

$$L_{\mathbf{X},\mathbf{d}} = \{\mathbf{w} : \mathbf{X}\mathbf{w} = \mathbf{d}\}.$$

A.2 Comparison of Performance Errors. Given weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$, a matrix $\mathbf{X}$ of input vectors, and a vector $\mathbf{d}$ of desired outputs, define

$$\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}, \mathbf{d}) = E_{pre} - E_{post},$$

where $E_{pre} = E(\mathbf{X}; \mathbf{w}_1, \mathbf{d})$ and $E_{post} = E(\mathbf{X}; \mathbf{w}_2, \mathbf{d})$. Let $\tilde{\mathbf{w}}$ be any element of $L_{\mathbf{X},\mathbf{d}}$. Then

$$\begin{aligned}
\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}, \mathbf{d}) &= \|\mathbf{X}\mathbf{w}_1 - \mathbf{d}\|^2 - \|\mathbf{X}\mathbf{w}_2 - \mathbf{d}\|^2 \\
&= \|\mathbf{X}\mathbf{w}_1\|^2 - \|\mathbf{X}\mathbf{w}_2\|^2 - 2\left(\mathbf{w}_1 - \mathbf{w}_2\right)^T \mathbf{X}^T\mathbf{d} \\
&= \left(\mathbf{w}_1 - \mathbf{w}_2\right)^T \mathbf{X}^T\mathbf{X}\left(\mathbf{w}_1 + \mathbf{w}_2\right) - 2\left(\mathbf{w}_1 - \mathbf{w}_2\right)^T \mathbf{X}^T\mathbf{X}\tilde{\mathbf{w}} \\
&= \left(\mathbf{w}_1 - \mathbf{w}_2\right)^T \mathbf{X}^T\mathbf{X}\left(\mathbf{w}_1 + \mathbf{w}_2 - 2\tilde{\mathbf{w}}\right).
\end{aligned} \tag{A.4}$$

Suppose given $n_i \times n$ matrices $\mathbf{X}_i$ and n_i-dimensional vectors $\mathbf{d}_i$ (for $i = 1, 2$). Put

$$L_i = L_{\mathbf{X}_i,\mathbf{d}_i} \quad \text{for } i = 1, 2.$$

If $\mathbf{X}_i$ has rank n_i, then

$$\mathbf{X}_i = \mathbf{T}_i\mathbf{Z}_i$$

for unique $n_i \times n_i$ and $n_i \times n$ matrices $\mathbf{T}_i$ and $\mathbf{Z}_i$ with $\mathbf{T}_i$ upper triangular and $\mathbf{Z}_i\mathbf{Z}_i^T = \mathbf{I}_{n_i}$. Note that the matrix $\mathbf{Z}_i^T\mathbf{Z}_i$ represents the operator that projects onto the image of $\mathbf{X}_i^T\mathbf{X}_i$, and so

$$\mathbf{Z}_i^T\mathbf{Z}_i\mathbf{X}_i^T\mathbf{X}_i = \mathbf{X}_i^T\mathbf{X}_i. \tag{A.5}$$

Let $\mathbf{w}_0$ be an element of $L_{\mathbf{X},\mathbf{d}}$, where

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \qquad \mathbf{d} = \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{pmatrix},$$

that is,

$$\left(\mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_2^T\mathbf{X}_2\right)\mathbf{w}_0 = \mathbf{X}_1^T\mathbf{d}_1 + \mathbf{X}_2^T\mathbf{d}_2. \tag{A.6}$$

(By equation A.3, such a $\mathbf{w}_0$ always exists.) Given $\mathbf{v}$ in $\mathbb{R}^n$, put

$$\mathbf{w}_1 = \mathbf{w}_0 + \mathbf{v}.$$

Let $\mathbf{w}_{02}$ and $\mathbf{w}_2$ be the orthogonal projections of $\mathbf{w}_0$ and $\mathbf{w}_1$, respectively, onto L_2. Then

$$\begin{aligned} \mathbf{X}_2^T\mathbf{X}_2\mathbf{w}_{02} &= \mathbf{X}_2^T\mathbf{d}_2 \qquad (A.7)\\ \mathbf{w}_2 &= \mathbf{w}_{02} + \left(\mathbf{I}_n - \mathbf{Z}_2^T\mathbf{Z}_2\right)(\mathbf{w}_1 - \mathbf{w}_{02}). \end{aligned}$$

Manipulation gives

$$\mathbf{w}_1 - \mathbf{w}_2 = \mathbf{Z}_2^T\mathbf{Z}_2(\mathbf{v} + \mathbf{w}_0 - \mathbf{w}_{02}), \tag{A.8}$$

and so

$$\mathbf{w}_1 + \mathbf{w}_2 - 2\mathbf{w}_0 = \left(2\mathbf{I}_n - \mathbf{Z}_2^T\mathbf{Z}_2\right)\mathbf{v} - \mathbf{Z}_2^T\mathbf{Z}_2(\mathbf{w}_0 - \mathbf{w}_{02}). \tag{A.9}$$

Let $\tilde{\mathbf{w}}$ be any element of $L_{\mathbf{X}_1,\mathbf{d}_1}$. Then equations A.4, A.6, A.7 to A.9, and A.5 yield

$$\begin{aligned} &\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1)\\ &\quad= (\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{X}_1^T\mathbf{X}_1(\mathbf{w}_1 + \mathbf{w}_2 - 2\tilde{\mathbf{w}})\\ &\quad= (\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{X}_1^T\mathbf{X}_1(\mathbf{w}_1 + \mathbf{w}_2) - 2(\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{X}_1^T\mathbf{d}_1\\ &\quad= (\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{X}_1^T\mathbf{X}_1(\mathbf{w}_1 + \mathbf{w}_2 - 2\mathbf{w}_0) - 2(\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{X}_2^T\mathbf{X}_2(\mathbf{w}_0 - \mathbf{w}_{02})\\ &\quad= (\mathbf{v} + \mathbf{w}_0 - \mathbf{w}_{02})^T\mathbf{Z}_2^T\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1(\mathbf{w}_1 + \mathbf{w}_2 - 2\mathbf{w}_0)\\ &\qquad- 2(\mathbf{v} + \mathbf{w}_0 - \mathbf{w}_{02})^T\mathbf{Z}_2^T\mathbf{Z}_2\mathbf{X}_2^T\mathbf{X}_2(\mathbf{w}_0 - \mathbf{w}_{02})\\ &\quad= (\mathbf{v} + \mathbf{w}_0 - \mathbf{w}_{02})^T\mathbf{Z}_2^T\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\left(2\mathbf{I}_n - \mathbf{Z}_2^T\mathbf{Z}_2\right)\mathbf{v}\\ &\qquad- (\mathbf{v} + \mathbf{w}_0 - \mathbf{w}_{02})^T\mathbf{Z}_2^T\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_2^T\mathbf{Z}_2(\mathbf{w}_0 - \mathbf{w}_{02})\\ &\qquad- 2(\mathbf{v} + \mathbf{w}_0 - \mathbf{w}_{02})^T\mathbf{Z}_2^T\mathbf{Z}_2\mathbf{X}_2^T\mathbf{X}_2(\mathbf{w}_0 - \mathbf{w}_{02}) \end{aligned}$$

$$
\begin{aligned}
&= \mathbf{v}^T \mathbf{Z}_2^T \mathbf{Z}_2 \mathbf{X}_1^T \mathbf{X}_1 \left(2\mathbf{I}_n - \mathbf{Z}_2^T \mathbf{Z}_2\right) \mathbf{v} \\
&\quad - 2\left(\mathbf{w}_0 - \mathbf{w}_{02}\right)^T \mathbf{Z}_2^T \mathbf{Z}_2 \left\{\mathbf{X}_1^T \mathbf{X}_1 \left(\mathbf{I}_n - \mathbf{Z}_2^T \mathbf{Z}_2\right) - \mathbf{X}_2^T \mathbf{X}_2\right\} \mathbf{v} \\
&\quad - \left(\mathbf{w}_0 - \mathbf{w}_{02}\right)^T \mathbf{Z}_2^T \mathbf{Z}_2 \left(2\mathbf{X}_2^T \mathbf{X}_2 + \mathbf{X}_1^T \mathbf{X}_1 \mathbf{Z}_2^T \mathbf{Z}_2\right) \left(\mathbf{w}_0 - \mathbf{w}_{02}\right) \\
&= \mathbf{v}^T \mathbf{Z}_2^T \mathbf{Z}_2 \mathbf{X}_1^T \mathbf{X}_1 \left(2\mathbf{I}_n - \mathbf{Z}_2^T \mathbf{Z}_2\right) \mathbf{v} \\
&\quad - 2\left(\mathbf{w}_0 - \mathbf{w}_{02}\right)^T \left\{\mathbf{Z}_2^T \mathbf{Z}_2 \mathbf{X}_1^T \mathbf{X}_1 \left(\mathbf{I}_n - \mathbf{Z}_2^T \mathbf{Z}_2\right) - \mathbf{X}_2^T \mathbf{X}_2\right\} \mathbf{v} \\
&\quad - \left(\mathbf{w}_0 - \mathbf{w}_{02}\right)^T \left(2\mathbf{X}_2^T \mathbf{X}_2 + \mathbf{Z}_2^T \mathbf{Z}_2 \mathbf{X}_1^T \mathbf{X}_1 \mathbf{Z}_2^T \mathbf{Z}_2\right) \left(\mathbf{w}_0 - \mathbf{w}_{02}\right). \qquad \text{(A.10)}
\end{aligned}
$$

A.3 Moments of Isotropic Distributions. In order to obtain results on the distribution of performance error, it is useful to have some moments of isotropic distributions.

Let $\mathbf{u}$ be uniformly distributed on S^{n-1}, and let $\mathbf{A}$ and $\mathbf{B}$ be $n \times n$ matrices. The formulas for the second and fourth moments of $\mathbf{u}$ given in equations 9.6.1 and 9.6.2 of Mardia and Jupp (2000), together with some algebraic manipulation, yield

$$
\mathrm{E}\left[\mathbf{u}^T \mathbf{A} \mathbf{u}\right] = \frac{\mathrm{tr}\left(\mathbf{A}\right)}{n} \qquad \text{(A.11)}
$$

$$
\mathrm{E}\left[\mathbf{u}^T \mathbf{A} \mathbf{u} \mathbf{u}^T \mathbf{B} \mathbf{u}\right] = \frac{\mathrm{tr}\left(\mathbf{AB}\right) + \mathrm{tr}\left(\mathbf{AB}^T\right) + \mathrm{tr}\left(\mathbf{A}\right)\mathrm{tr}\left(\mathbf{B}\right)}{n(n+2)} \qquad \text{(A.12)}
$$

$$
\mathrm{var}\left(\mathbf{u}^T \mathbf{A} \mathbf{u}\right) = \frac{n\mathrm{tr}\left(\mathbf{A}^2\right) + n\mathrm{tr}\left(\mathbf{A}\mathbf{A}^T\right) - 2\mathrm{tr}\left(\mathbf{A}\right)^2}{n^2(n+2)}. \qquad \text{(A.13)}
$$

Now let $\mathbf{x}$ be isotropically distributed on $\mathbb{R}^n$, that is, $\mathbf{U}\mathbf{x}$ has the same distribution as $\mathbf{x}$ for all orthogonal $n \times n$ matrices $\mathbf{U}$. Then writing $\mathbf{x} = \|\mathbf{x}\|\mathbf{u}$ with $\|\mathbf{u}\| = 1$ and using equations A.11 to A.13 gives

$$
\mathrm{E}\left[\mathbf{x}^T \mathbf{A} \mathbf{x}\right] = \frac{\mathrm{E}\left[\|\mathbf{x}\|^2\right]\mathrm{tr}\left(\mathbf{A}\right)}{n} \qquad \text{(A.14)}
$$

$$
\mathrm{E}\left[\mathbf{x}^T \mathbf{A} \mathbf{x} \mathbf{x}^T \mathbf{B} \mathbf{x}\right] = \frac{\mathrm{E}\left[\|\mathbf{x}\|^4\right]\left\{\mathrm{tr}\left(\mathbf{AB}\right) + \mathrm{tr}\left(\mathbf{AB}^T\right) + \mathrm{tr}\left(\mathbf{A}\right)\mathrm{tr}\left(\mathbf{B}\right)\right\}}{n(n+2)}
$$

$$
\begin{aligned}
\mathrm{var}\left(\mathbf{x}^T \mathbf{A} \mathbf{x}\right) &= \frac{\mathrm{E}\left[\|\mathbf{x}\|^4\right]\left\{n\mathrm{tr}\left(\mathbf{A}^2\right) + n\mathrm{tr}\left(\mathbf{A}\mathbf{A}^T\right) - 2\mathrm{tr}\left(\mathbf{A}\right)^2\right\}}{n^2(n+2)} \\
&\quad + \frac{\mathrm{var}\left(\|\mathbf{x}\|^2\right)\mathrm{tr}\left(\mathbf{A}\right)^2}{n^2}. \qquad \text{(A.15)}
\end{aligned}
$$

A.4 Distribution of Performance Error. Now suppose that $\mathbf{X}_1$, $\mathbf{d}_1$, $\mathbf{X}_2$, $\mathbf{d}_2$, and $\mathbf{v}$ are random and satisfy

$$\begin{aligned} &\mathbf{X}_1 \text{ and } \mathbf{v} \text{ are independent,} \\ &\text{the distribution of } \mathbf{X}_1 \text{is isotropic,} \\ &\mathbf{v} \text{ has an isotropic distribution,} \end{aligned} \tag{A.16}$$

where conditions A.16 mean that $\mathbf{U}\mathbf{X}_1\mathbf{V}$ has the same distribution as $\mathbf{X}_1$ for all orthogonal $n_1 \times n_1$ matrices $\mathbf{U}$ and all orthogonal $n \times n$ matrices $\mathbf{V}$. Then equation A.10 yields

$$\begin{aligned} &\mathrm{E}\left[\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \,|\mathbf{X}_1, \mathbf{X}_2\right] \\ &\quad = \frac{\mathrm{E}\left[\|\mathbf{v}\|^2\right]}{n} \mathrm{tr}\left(\mathbf{X}_1^T \mathbf{X}_1 \mathbf{Z}_2^T \mathbf{Z}_2\right) \\ &\qquad - \left(\mathbf{w}_0 - \mathbf{w}_{02}\right)^T \left(2\mathbf{X}_2^T \mathbf{X}_2 + \mathbf{Z}_2^T \mathbf{Z}_2 \mathbf{X}_1^T \mathbf{X}_1 \mathbf{Z}_2^T \mathbf{Z}_2\right) \left(\mathbf{w}_0 - \mathbf{w}_{02}\right). \end{aligned} \tag{A.17}$$

Taking expectations over $\mathbf{X}_1$ and $\mathbf{X}_2$ in equation A.17 gives the following general result on FLL:

$$\begin{aligned} &\mathrm{E}[\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1)] > 0 \text{ iff} \\ &\qquad \mathrm{E}[\|\mathbf{v}\|^2] > \frac{n^2 \mathrm{E}\left[(\mathbf{w}_0 - \mathbf{w}_{02})^T \left(2\mathbf{X}_2^T \mathbf{X}_2 + \mathbf{Z}_2^T \mathbf{Z}_2 \mathbf{X}_1^T \mathbf{X}_1 \mathbf{Z}_2^T \mathbf{Z}_2\right)(\mathbf{w}_0 - \mathbf{w}_{02})\right]}{n_1 n_2}. \end{aligned} \tag{A.18}$$

The intuitive interpretation of this result is that if $\mathrm{E}\left[\|\mathbf{v}\|^2\right]$ is large enough, then there is FLL, whereas if $P\left(\mathbf{w}_0 \neq \mathbf{w}_{02}\right) > 0$ then "negative FLL" can occur. In particular, if $n_1 + n_2 \leq n$ and $P\left(\mathbf{v} \neq \mathbf{0}\right) > 0$, then there is FLL.

A.5 The Case $n_1 + n_2 \leq n$. In this section we assume that $\mathbf{X}_1$, $\mathbf{d}_1$, $\mathbf{X}_2$ and $\mathbf{d}_2$ are random and that

$$(\mathbf{X}_1, \mathbf{d}_1), (\mathbf{X}_2, \mathbf{d}_2) \text{ and } \mathbf{v} \text{ are independent,} \tag{A.19}$$

$$\text{the distribution of } \mathbf{v} \text{ is isotropic.} \tag{A.20}$$

We suppose also that $n_1 + n_2 \leq n$, and that the distributions of $\mathbf{X}_1$, $\mathbf{d}_1$, $\mathbf{X}_2$, and $\mathbf{d}_2$ are continuous. Then, with probability 1,

$$\mathbf{X}_1 \mathbf{w}_0 = \mathbf{d}_1 \text{ and } \mathbf{X}_2 \mathbf{w}_0 = \mathbf{d}_2,$$

so that $\mathbf{w}_{02} = \mathbf{w}_0$ and equation A.10 reduces to

$$\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) = \mathbf{v}^T \mathbf{Z}_2^T \mathbf{Z}_2 \mathbf{X}_1^T \mathbf{X}_1 \left(2\mathbf{I}_n - \mathbf{Z}_2^T \mathbf{Z}_2\right) \mathbf{v}. \tag{A.21}$$

A.5.1 FLL Is More Probable Than Not. Let $\mathbf{w}_1^*$ be the reflection of $\mathbf{w}_1$ in L_2, that is,

$$\mathbf{w}_1^* = \mathbf{w}_2 - (\mathbf{w}_1 - \mathbf{w}_2)\,.$$

Consideration of the parallelogram with vertices at $\mathbf{w}_0$, $\mathbf{w}_1$, $\mathbf{w}_1^*$, and $\mathbf{w}_1 + \mathbf{w}_1^* - \mathbf{w}_0$ gives

$$\begin{aligned}
&2\left(\|\mathbf{X}_1(\mathbf{w}_1 - \mathbf{w}_0)\|^2 + \|\mathbf{X}_1(\mathbf{w}_1^* - \mathbf{w}_0)\|^2\right) \\
&\quad = \|\mathbf{X}_1\left([\mathbf{w}_1 - \mathbf{w}_0] + \left[\mathbf{w}_1^* - \mathbf{w}_0\right]\right)\|^2 + \|\mathbf{X}_1\left([\mathbf{w}_1 - \mathbf{w}_0] - \left[\mathbf{w}_1^* - \mathbf{w}_0\right]\right)\|^2 \\
&\quad = 4\left(\|\mathbf{X}_1(\mathbf{w}_2 - \mathbf{w}_0)\|^2 + \|\mathbf{X}_1(\mathbf{w}_1 - \mathbf{w}_2)\|^2\right),
\end{aligned}$$

so that (since $\mathbf{d}_1 = \mathbf{X}_1\mathbf{w}_0$)

$$\begin{aligned}
&\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) + \delta(\mathbf{w}_1^*, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \\
&\quad = \|\mathbf{X}_1(\mathbf{w}_1 - \mathbf{w}_0)\|^2 + \|\mathbf{X}_1(\mathbf{w}_1^* - \mathbf{w}_0)\|^2 - 2\|\mathbf{X}_1(\mathbf{w}_2 - \mathbf{w}_0)\|^2 \\
&\quad = 2\|\mathbf{X}_1(\mathbf{w}_1 - \mathbf{w}_2)\|^2 \geq 0.
\end{aligned}$$

Thus if $\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) < 0$, then $\delta(\mathbf{w}_1^*, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0$. If $\mathbf{v}$ is distributed isotropically, then $\mathbf{w}_1^* - \mathbf{w}_0$ is distributed isotropically, so that $\delta(\mathbf{w}_1^*, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1)$ has the same distribution (conditionally on $\mathbf{X}_1$, $\mathbf{d}_1$ and $\mathbf{X}_2$) as $\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1)$, and so

$$\begin{aligned}
P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) < 0 | \mathbf{X}_1, \mathbf{d}_1, \mathbf{X}_2) &\leq P(\delta(\mathbf{w}_1^*, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0 | \mathbf{X}_1, \mathbf{d}_1, \mathbf{X}_2) \\
&= P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0 | \mathbf{X}_1, \mathbf{d}_1, \mathbf{X}_2).
\end{aligned} \tag{A.22}$$

Further, if $\mathbf{v} \in L_2 \setminus L_1$, then $\mathbf{w}_2 = \mathbf{w}_1 = \mathbf{w}_1^*$, so that $\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) = \delta(\mathbf{w}_1^*, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0$. By continuity of δ, there is a neighborhood of $\mathbf{v}$ on which $\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0$ and $\delta(\mathbf{w}_1^*, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0$. Thus, if $L_2 \setminus L_1 \neq \emptyset$, then equation A.22 can be refined to

$$P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) < 0 | \mathbf{X}_1, \mathbf{d}_1, \mathbf{X}_2) < P(\delta(\mathbf{w}_1^*, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) < 0 | \mathbf{X}_1, \mathbf{d}_1, \mathbf{X}_2). \tag{A.23}$$

Since $P(L_2 \subset L_1) = 0$ and $P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) < 0 | \mathbf{X}_1, \mathbf{d}_1, \mathbf{X}_2)$ is a continuous function of $\mathbf{X}_1$, $\mathbf{d}_1$ and $\mathbf{X}_2$, it follows from equation A.23 that

$$P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) < 0) < P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0),$$

which implies the following result.

Theorem 1

$$P(\delta(\boldsymbol{w}_1, \boldsymbol{w}_2; \boldsymbol{X}_1, \boldsymbol{d}_1) > 0) > 0.5.$$

This implies that the median of $\delta(\boldsymbol{w}_1, \boldsymbol{w}_2; \boldsymbol{X}_1, \boldsymbol{d}_1)$ is positive.

A.5.2 A Lower Bound for $P(\delta > 0)$. Our proof depends on Chebyshev's inequality, which states that for any positive value of t,

$$P(|\delta - \mathrm{E}[\delta]| \geq t) \leq \frac{\mathrm{var}(\delta)}{t^2},$$

where $\mathrm{var}(\delta)$ denotes the variance of δ. If we set $t = \mathrm{E}[\delta]$, then (since, by equation A.28, $\mathrm{E}[\delta] > 0$)

$$P(\delta \leq 0) \leq \frac{\mathrm{var}(\delta)}{\mathrm{E}[\delta]^2}. \tag{A.24}$$

This provides a lower bound for the probability of FLL. We prove that this bound approaches unity as n approaches infinity.

Now we assume (in addition to conditions A.19 and A.20) that

$$\text{the distributions of } \mathbf{X}_1 \text{ and } \mathbf{X}_2 \text{ are isotropic.} \tag{A.25}$$

It follows from equations A.21, A.14, and A.15 that

$$\begin{aligned}\mathrm{E}[\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) | \mathbf{Z}_2, \mathbf{v}] &= \mathbf{v}^T \mathbf{Z}_2^T \mathbf{Z}_2 \mathrm{E}\left[\|\mathbf{x}\|^2\right] \frac{n_1}{n} \mathbf{I}_n \left(2\mathbf{I}_n - \mathbf{Z}_2^T \mathbf{Z}_2\right) \mathbf{v} \\ &= \frac{n_1}{n} \mathrm{E}\left[\|\mathbf{x}\|^2\right] \mathbf{v}^T \mathbf{Z}_2^T \mathbf{Z}_2 \mathbf{v}, \end{aligned} \tag{A.26}$$

where $\mathbf{x}$ is the first column of $\mathbf{X}_1^T$, and

$$\begin{aligned}&\mathrm{var}(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) | \mathbf{Z}_2, \mathbf{v}) \\ &\quad = n_1 \left\{ \frac{\mathrm{E}\left[\|\mathbf{x}\|^4\right] \left\{(n-2)\|\mathbf{Z}_2\mathbf{v}\|^4 + n\|\mathbf{Z}_2\mathbf{v}\|^2 \| \left(2\mathbf{I}_n - \mathbf{Z}_2^T \mathbf{Z}_2\right) \mathbf{v}\|^2\right\}}{n^2(n+2)} \right. \\ &\qquad \left. + \frac{\mathrm{var}\left(\|\mathbf{x}\|^2\right) \|\mathbf{Z}_2\mathbf{v}\|^4}{n^2} \right\}. \end{aligned} \tag{A.27}$$

Since $\mathbf{v}$ has an isotropic distribution, equations A.26, A.11, and A.13 imply that

$$\mathrm{E}[\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) | \mathbf{Z}_2, \|\mathbf{v}\|] = \frac{n_1 n_2}{n^2} \mathrm{E}\left[\|\mathbf{x}\|^2\right] \|\mathbf{v}\|^2. \tag{A.28}$$

Given that there are n_1 associations in the subset A_1 that is not relearned, equation A.28 implies the following theorem about the expected amount of recovery per association in A_1.

Theorem 2

$$E\left[\frac{\delta(\boldsymbol{w}_1, \boldsymbol{w}_2; \boldsymbol{X}_1, \boldsymbol{d}_1)}{n_1}\middle| \mathbf{Z}_2, \|\boldsymbol{v}\|\right] = \frac{n_2}{n^2} E[\|\boldsymbol{x}\|^2]\|\boldsymbol{v}\|^2. \tag{A.29}$$

Equations A.26 and A.13 also imply that

$$\begin{aligned}
&\mathrm{var}\left(\mathrm{E}\left[\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1)\,|\mathbf{Z}_2, \mathbf{v}\right]|\mathbf{Z}_2, \|\mathbf{v}\|\right) \\
&\quad = \left(\frac{n_1}{n}\mathrm{E}\left[\|\mathbf{x}\|^2\right]\right)^2 \frac{\|\mathbf{v}\|^4\left(2nn_2 - 2n_2^2\right)}{n^2(n+2)} \\
&\quad = \frac{2n_1^2 n_2(n-n_2)\mathrm{E}\left[\|\mathbf{x}\|^2\right]^2\|\mathbf{v}\|^4}{n^4(n+2)},
\end{aligned} \tag{A.30}$$

and it follows from equations A.27 and A.12 that

$$\begin{aligned}
&\mathrm{E}\left[\mathrm{var}\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1)\,|\mathbf{Z}_2, \mathbf{v}\right)|\mathbf{Z}_2, \|\mathbf{v}\|\right] \\
&\quad = \frac{n_1\|\mathbf{v}\|^3}{n(n+2)}\left\{\frac{\mathrm{E}\left[\|\mathbf{x}\|^4\right]\left\{(n-2)n_2(n_2+2) + nn_2(2n-n_2+2)\right\}}{n^2(n+2)}\right. \\
&\qquad \left. + \frac{\mathrm{var}\left(\|\mathbf{x}\|^2\right)n_2(n_2+2)}{n^2}\right\} \\
&\quad = \frac{n_1 n_2\|\mathbf{v}\|^4}{n^3(n+2)^2}\left\{\mathrm{E}\left[\|\mathbf{x}\|^4\right]2(n^2+2n-n_2-2) + \mathrm{var}\left(\|\mathbf{x}\|^2\right)(n+2)(n_2+2)\right\}.
\end{aligned} \tag{A.31}$$

Then equations A.30 and A.31 give

$$\begin{aligned}
&\mathrm{var}\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1)\,|\mathbf{Z}_2, \|\mathbf{v}\|\right) \\
&\quad = \frac{2n_1^2 n_2(n-n_2)\mathrm{E}\left[\|\mathbf{x}\|^2\right]^2\|\mathbf{v}\|^4}{n^4(n+2)} \\
&\qquad + \frac{n_1 n_2\|\mathbf{v}\|^4}{n^3(n+2)^2}\left\{\mathrm{E}\left[\|\mathbf{x}\|^4\right]2(n^2+2n-n_2-2) + \mathrm{var}\left(\|\mathbf{x}\|^2\right)(n+2)(n_2+2)\right\}
\end{aligned}$$

$$= \frac{n_1 n_2 \|\mathbf{v}\|^4}{n^4(n+2)^2}\{2[n_1(n+2)(n-n_2)+n(n-n_2)+n(n+2)(n-1)]\mathrm{E}[\|\mathbf{x}\|^2]^2$$
$$+ n^2(2n+n_2+6)\mathrm{var}(\|\mathbf{x}\|^2)\},$$

and so

$$\frac{\mathrm{var}\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1)\,|\mathbf{Z}_2, \|\mathbf{v}\|\right)}{\mathrm{E}\left[\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1)\,|\mathbf{Z}_2, \|\mathbf{v}\|\right]^2} = \frac{a_0(n, n_1, n_2) + a_1(n, n_2)\gamma(n)}{n_1 n_2 (n+2)^2},$$

where

$$a_0(n, n_1, n_2) = 2\{n_1(n+2)(n-n_2) + n(n-n_2) + n(n+2)(n-1)\}$$
$$a_1(n, n_2) = n^2(2n + n_2 + 6)$$
$$\gamma(n) = \frac{\mathrm{var}\left(\|\mathbf{x}\|^2\right)}{\mathrm{E}\left[\|\mathbf{x}\|^2\right]^2}.$$

Chebyshev's inequality implies the following theorem.

Theorem 3

$$P\left(\delta(\boldsymbol{w}_1, \boldsymbol{w}_2; \boldsymbol{X}_1, \boldsymbol{d}_1) \le 0\,|\boldsymbol{Z}_2, \|\boldsymbol{v}\|\right) \le \frac{a_0(n, n_1, n_2) + a_1(n, n_1, n_2)\gamma(n)}{n_1 n_2 (n+2)^2}.$$

Since the right-hand side does not depend on $\mathbf{Z}_2$ or $\|\mathbf{v}\|$, this gives the following result.

If $\gamma(n)/n \to 0$ and $n_1/n, n_2/n$ are bounded away from zero as $n \to \infty$, then

$$P\left(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0\right) \to 1, \qquad n \to \infty.$$

Example. If

$$\mathbf{x} \sim N\left(\mathbf{0}, \sigma_x^2 \mathbf{I}_n\right),$$

then

$$\mathrm{E}\left[\|\mathbf{x}\|^2\right] = n\sigma_x^2, \qquad \mathrm{var}\left(\|\mathbf{x}\|^2\right) = 2n\sigma_x^4, \qquad \gamma(n) = \frac{2}{n},$$

and so

$$P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) > 0) \to 1, \qquad n \to \infty,$$

provided that n_1/n and n_2/n are bounded away from zero.

A.5.3 Learning A_3 Instead of A_2. Now suppose that relearning of A_2 is replaced by learning another subset A_3 of n_2 associations. Let the matrix $\mathbf{X}_3$ and vector $\mathbf{d}_3$ be such that the subspace L_3 corresponding to A_3 has the form $L_3 = L_{\mathbf{X}_3,\mathbf{d}_3}$.

Let $\mathbf{w}_3$ and $\mathbf{w}_{13}$ denote the orthogonal projections of $\mathbf{w}_1$ onto L_3 and $L_1 \cap L_3$, respectively. Then

$$\mathbf{w}_3 = \mathbf{w}_{13} + \left(\mathbf{I}_n - \mathbf{Z}_3^T\mathbf{Z}_3\right)(\mathbf{w}_1 - \mathbf{w}_{13}), \tag{A.32}$$

and so

$$\mathbf{w}_1 = \mathbf{w}_3 + \mathbf{Z}_3^T\mathbf{Z}_3(\mathbf{w}_1 - \mathbf{w}_{13}). \tag{A.33}$$

From equation A.4 with $\tilde{\mathbf{w}} = \mathbf{w}_{13}$, and equations A.33 and A.32, we have

$$\begin{aligned}\delta(\mathbf{w}_1, \mathbf{w}_3; \mathbf{X}_1, \mathbf{d}_1) &= (\mathbf{w}_1 - \mathbf{w}_3)^T \mathbf{X}_1^T\mathbf{X}_1 (\mathbf{w}_1 + \mathbf{w}_3 - 2\mathbf{w}_{13}) \\ &= (\mathbf{w}_1 - \mathbf{w}_{13})^T \mathbf{Z}_3^T\mathbf{Z}_3\mathbf{X}_1^T\mathbf{X}_1 (\mathbf{w}_1 + \mathbf{w}_3 - 2\mathbf{w}_{13}) \\ &= (\mathbf{v} - \tilde{\boldsymbol{\omega}})^T \mathbf{Z}_3^T\mathbf{Z}_3\mathbf{X}_1^T\mathbf{X}_1 \left(2\mathbf{I}_n - \mathbf{Z}_3^T\mathbf{Z}_3\right)(\mathbf{v} - \tilde{\boldsymbol{\omega}}), \end{aligned} \tag{A.34}$$

where

$$\tilde{\boldsymbol{\omega}} = \mathbf{w}_{13} - \mathbf{w}_0.$$

Since $\mathbf{X}_1\mathbf{w}_0 = \mathbf{X}_1\mathbf{w}_{13}$, equation A.34 can be expanded as

$$\begin{aligned}&\delta(\mathbf{w}_1, \mathbf{w}_3; \mathbf{X}_1, \mathbf{d}_1) \\ &\quad= \mathbf{v}^T\mathbf{Z}_3^T\mathbf{Z}_3\mathbf{X}_1^T\mathbf{X}_1 \left(2\mathbf{I}_n - \mathbf{Z}_3^T\mathbf{Z}_3\right)\mathbf{v} \\ &\qquad- \mathbf{v}^T\mathbf{Z}_3^T\mathbf{Z}_3\mathbf{X}_1^T\mathbf{X}_1 \left(2\mathbf{I}_n - \mathbf{Z}_3^T\mathbf{Z}_3\right)\tilde{\boldsymbol{\omega}} - \tilde{\boldsymbol{\omega}}^T\mathbf{Z}_3^T\mathbf{Z}_3\mathbf{X}_1^T\mathbf{X}_1 \left(2\mathbf{I}_n - \mathbf{Z}_3^T\mathbf{Z}_3\right)\mathbf{v} \\ &\qquad- \tilde{\boldsymbol{\omega}}^T\mathbf{Z}_3^T\mathbf{Z}_3\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_3^T\mathbf{Z}_3\tilde{\boldsymbol{\omega}},\end{aligned}$$

and so

$$\begin{aligned}&\mathrm{E}\left[\delta(\mathbf{w}_1, \mathbf{w}_3; \mathbf{X}_1, \mathbf{d}_1) | \mathbf{X}_1, \mathbf{d}_1, \mathbf{X}_2, \mathbf{d}_2, \mathbf{X}_3, \mathbf{d}_3\right] \\ &\quad= \frac{\mathrm{E}\left[\|\mathbf{v}\|^2\right]}{n} \mathrm{tr}\left(\mathbf{Z}_3^T\mathbf{Z}_3\mathbf{X}_1^T\mathbf{X}_1 \left(2\mathbf{I}_n - \mathbf{Z}_3^T\mathbf{Z}_3\right)\right) - \tilde{\boldsymbol{\omega}}^T\mathbf{Z}_3^T\mathbf{Z}_3\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_3^T\mathbf{Z}_3\tilde{\boldsymbol{\omega}} \\ &\quad= \frac{\mathrm{E}\left[\|\mathbf{v}\|^2\right]}{n} \mathrm{tr}\left(\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_3^T\mathbf{Z}_3\right) - \|\mathbf{X}_1\mathbf{Z}_3^T\mathbf{Z}_3\tilde{\boldsymbol{\omega}}\|^2.\end{aligned}$$

Now assume that

$(\mathbf{X}_1, \mathbf{d}_1)$, $(\mathbf{X}_2, \mathbf{d}_2)$, $(\mathbf{X}_3, \mathbf{d}_3)$ and $\mathbf{v}$ are independent,

the distributions of $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ and $\mathbf{v}$ are isotropic.

Since

$$\frac{\mathrm{E}\left[\|\mathbf{v}\|^2\right]}{n}\mathrm{E}\left[\mathrm{tr}\left(\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_2^T\mathbf{Z}_2\right)\right] = \frac{\mathrm{E}\left[\|\mathbf{v}\|^2\right]}{n}\mathrm{E}\left[\mathrm{tr}\left(\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_3^T\mathbf{Z}_3\right)\right]$$
$$= \mathrm{E}\left[\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1)\right],$$

we have the following theorem.

Theorem 4

$$E[\delta(\boldsymbol{w}_1, \boldsymbol{w}_3; \boldsymbol{X}_1, \boldsymbol{d}_1)] \leq E\left[\delta(\boldsymbol{w}_1, \boldsymbol{w}_2; \boldsymbol{X}_1, \boldsymbol{d}_1)\right].$$

Appendix B: Behavior of the Gradient Algorithm

If E is regarded as a function of $\mathbf{w}$, then differentiation of equation A.2 shows that the gradient of E at $\mathbf{w}$ is

$$\nabla E_{(\mathbf{w})} = 2\mathbf{X}^T\left(\mathbf{X}\mathbf{w} - \mathbf{d}\right).$$

Then for any algorithm that takes an initial $\mathbf{w}^{(0)}$ to $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ using steps $\mathbf{w}^{(t+1)} - \mathbf{w}^{(t)}$ in the direction of $\nabla E_{(\mathbf{w}^{(t)})}$, $\mathbf{w}^{(t)} - \mathbf{w}^{(0)}$ is in the image of $\mathbf{X}^T\mathbf{X}$, and so is orthogonal to $L_{\mathbf{X},\mathbf{d}}$. It follows that if $\|\mathbf{X}\mathbf{w}^{(t)} - \mathbf{d}\|^2 \to \min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{d}\|^2$ as $t \to \infty$, then $\mathbf{w}^{(t)}$ converges to the orthogonal projection of $\mathbf{w}^{(0)}$ onto $L_{\mathbf{X},\mathbf{d}}$.

Appendix C: The Geometry of Performance Error When $n_1 = 1$

Given associations A_1 and A_2, we prove that if $n_1 = 1$ and input vectors have unit length (so that $\|\mathbf{x}_1\| = 1$), then the difference δ in performance errors on association A_1 of $\mathbf{w}_1$ (i.e., after partial forgetting) and $\mathbf{w}_2$ (i.e., after relearning A_2) is equal to the difference $Q = p^2 - q^2$. This proof supports the geometric account given in the article and in Figure 2 and does not (in general) apply if $n_1 > 1$.

We begin by proving that (if $n_1 = 1$ and $\|\mathbf{x}_1\| = 1$) the performance error of an association A_1 for an arbitrary weight vector $\mathbf{w}_1$ is equal to the squared distance p^2 between $\mathbf{w}_1$ and its orthogonal projection $\mathbf{w}_1'$ onto the affine subspace L_1 corresponding to A_1. If $n_1 = 1$, then L_1 has the form

$$L_1 = \{\mathbf{w} : \mathbf{w} \cdot \mathbf{x}_1 = d_1\}$$

for some $\mathbf{x}_1$ and d_1. Given an arbitrary weight vector $\mathbf{w}_1$, we define the performance error on association A_1 as equivalent to

$$E(\mathbf{w}_1, A_1) = (\mathbf{w}_1 \cdot \mathbf{x}_1 - d_1)^2. \tag{C.1}$$

The orthogonal projection $\mathbf{w}_1'$ of $\mathbf{w}_1$ onto L_1 is

$$\mathbf{w}_1' = \mathbf{w}_1 + \frac{d_1 - \mathbf{w}_1 \cdot \mathbf{x}_1}{\|\mathbf{x}_1\|^2}\mathbf{x}_1, \tag{C.2}$$

so that

$$d_1 = \mathbf{w}_1' \cdot \mathbf{x}_1. \tag{C.3}$$

Substituting equation C.3 into C.1 and using C.2 yields

$$\begin{aligned} E(\mathbf{w}_1, A_1) &= \|\mathbf{w}_1 - \mathbf{w}_1'\|^2 \|\mathbf{x}_1\|^2 \\ &= p^2 \|\mathbf{x}_1\|^2. \end{aligned} \tag{C.4}$$

Now suppose that $\|\mathbf{x}_1\| = 1$. Then

$$E(\mathbf{w}_1, A_1) = p^2,$$

that is, the performance error is equal to the squared distance between the weight vectors $\mathbf{w}_1$ and $\mathbf{w}_1'$. The same line of reasoning can be applied to prove that

$$E(\mathbf{w}_2, A_1) = q^2.$$

Thus, the difference δ in performance error on A_1 for weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ is

$$\begin{aligned} \delta &= E(\mathbf{w}_1, A_1) - E(\mathbf{w}_2, A_1) \\ &= p^2 - q^2 \\ &= Q. \end{aligned}$$

Acknowledgments

Thanks to S. Isard for substantial help with the analysis presented here; to R. Lister, S. Eglen, P. Parpia, A. Farthing, P. Warren, K. Gurney, N. Hunkin, and two anonymous referees for comments; and J. Porrill for useful discussions.

References

Atkins, P. (2001). What happens when we relearn part of what we previously knew? Predictions and constraints for models of long-term memory. *Psychological Research, 65*(3), 202–215.

Atkins, P., & Murre, J., (1998). Recovery of unrehearsed items in connectionist models. *Connection Science, 10*(2), 99–119.

Coltheart, M., & Byng, S. (1989). A treatment for surface dyslexia. In X. Seron (Ed.), *Cognitive approaches in neuropsychological rehabilitation*. Mahwah, NJ: Erlbaum.

Dean, P., Porrill, J., & Stone, J. V. (2002). Decorrelation control by the cerebellum achieves oculomotor plant compensation in simulated vestibulo-ocular reflex. *Proceedings Royal Society (B), 269*(1503), 1895–1904.

Dienes, Z., Altmann, G., & Gao, S.-J. (1999). Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. *Cognitive Science, 23*, 53–82.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks, 12*(7–8), 961–974.

Goldin, M., Segal, M., & Avignone, E. (2001). Functional plasticity triggers formation and pruning of dendritic spines in cultured hippocampal networks. *J. Neuroscience, 21*(1), 186–193.

Hanson, S. J., & Negishi, M. (2002). On the emergence of rules in neural networks. *Neural Computation, 14*, 2245–2268.

Harvey, I., & Stone, J.V. (1996). Unicycling helps your French: Spontaneous recovery of associations by learning unrelated tasks. *Neural Computation, 8*, 697–704.

Hinton, G., & Plaut, D. (1987). Using fast weights to deblur old memories. In *Proceedings Ninth Annual Conference of the Cognitive Science Society*, Seattle WA, 177–186.

Mardia, K. V., & Jupp, P. E. (2000). *Directional statistics*. New York: Wiley.

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron, 41*(2), 269–280.

Purves D., & Lichtman, J. (1980). Elimination of synapses in the developing nervous system. *Science, 210*, 153–157.

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., & Frackowiak, R. (2004). Temporal difference models describe higher order learning in humans. *Nature, 429*, 664–667.

Stone, J. V., Hunkin, N. M., & Hornby, A. (2001). Predicting spontaneous recovery of memory. *Nature, 414*, 167–168.

Sutton, R. (1988). Learning to predict by the methods of temporal differences. *Machine Learning, 3*, 9–44.

Weekes, B., & Coltheart, M. (1996). Surface dyslexia and surface dysgraphia: Treatment studies and their theoretical implications. *Cognitive Neuropsychology, 13*, 277–315.

Received August 1, 2005; accepted May 19, 2006.