Falling towards Forgetfulness: Synaptic Decay Prevents Spontaneous Recovery of Memory

James V. Stone[1]*, Peter E. Jupp[2]

1 Psychology Department, Sheffield University, Sheffield, United Kingdom, **2** School of Mathematics and Statistics, University of St Andrews, St Andrews, United Kingdom

Abstract

Long after a new language has been learned and forgotten, relearning a few words seems to trigger the recall of other words. This "free-lunch learning" (FLL) effect has been demonstrated both in humans and in neural network models. Specifically, previous work proved that linear networks that learn a set of associations, then partially forget them all, and finally relearn some of the associations, show improved performance on the remaining (i.e., nonrelearned) associations. Here, we prove that relearning forgotten associations *decreases* performance on nonrelearned associations; an effect we call negative free-lunch learning. The difference between free-lunch learning and the negative free-lunch learning presented here is due to the particular method used to induce forgetting. Specifically, if forgetting is induced by isotropic *drifting* of weight vectors (i.e., by adding isotropic noise), then free-lunch learning is observed. However, as proved here, if forgetting is induced by weight values that simply decay or *fall* towards zero, then negative free-lunch learning is observed. From a biological perspective, and assuming that nervous systems are analogous to the networks used here, this suggests that evolution may have selected physiological mechanisms that involve forgetting using a form of synaptic drift rather than synaptic decay, because synaptic drift, but not synaptic decay, yields free-lunch learning.

Citation: Stone JV, Jupp PE (2008) Falling towards Forgetfulness: Synaptic Decay Prevents Spontaneous Recovery of Memory. PLoS Comput Biol 4(8): e1000143. doi:10.1371/journal.pcbi.1000143

Editor: Karl J. Friston, University College London, United Kingdom

Received December 7, 2007; **Accepted** June 25, 2008; **Published** August 22, 2008

Funding: No funding was received for this work.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: j.v.stone@sheffield.ac.uk

Introduction

The idea that structural changes underpin the formation of new memories can be traced to the 19th century [1]. More recently, Hebb proposed that "When an axon of cell A is near enough to excite B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased" [2]. It is now widely accepted that learning involves some form of Hebbian adaptation, and a growing body of evidence suggests that Hebbian adaptation is associated with the long-term potentiation (LTP) observed in neuronal systems [3]. LTP is an increase in synaptic efficacy which occurs in the presence of pre-synaptic and post-synaptic activity, and can be specific to a single synapse. One consequence of Hebbian adaptation is that information regarding a specific association is distributed amongst many synaptic connections, and therefore gives rise to a distributed representation of each association.

In [4], participants learned the layout of letters on a "scrambled" keyboard. After a period of forgetting, participants relearned a subset of letter positions. Crucially, this improved performance on the remaining (i.e., nonrelearned) letter positions. However, whereas relearning some associations shows evidence of FLL in some studies [4–6], this is not found in not all studies [7]. This discrepancy may be because the many studies performed to investigate this general phenomenon use a wide variety of different materials and procedures, with some measuring recall and others measuring recognition performance, for example. However, within the realms of psychology, one relevant effect is known as part-set cueing inhibition.

Part-set cueing inhibition [8] occurs when a subject is exposed to part of a set of previously learned items, which is found to reduce recall of nonrelearned items. However, [9] showed that a learned row of words was better recalled if the cues consisted of a subset of words placed in their learned positions than if cue words were placed in other positions. In this case, part-set cueing seems to improve performance, but only if each "part" appears in the spatial position in which it was originally learned. This position-specificity is consistent with the FLL effect reported using the "scrambled keyboard" procedure in [4] but has no obvious concomitant in network models (e.g., [4,10,11]).

If the brain stores information as distributed representations, then each neuron contributes to the storage of many associations. Therefore, relearning some old and partially forgotten associations should affect the integrity of other associations learned at about the same time. As noted above, previous work has shown that relearning some forgotten associations does not disrupt other associations, but partially restores them. This FLL effect has also been demonstrated in neural network models ([10,12]), where it can accelerate evolution of adaptive behaviors [13]. Crucially, in [12], the proof that relearning some associations partially restores other associations assumes that forgetting is caused by the addition of isotropic noise to connection weights, which could result from the cumulative effect of small random changes in connection weights. In contrast, here we prove that if forgetting is induced by shrinking weights towards zero, so that weights

Author Summary

If you learn a skill, then partially forget it, does relearning part of that skill induce recovery of other parts of the skill? More generally, if you learn a set of associations, then partially forget them, does relearning a subset induce recovery of the remaining associations? In previous work, in which participants learned the layout of a scrambled computer keyboard, the answer to this question appeared to be "yes." More recently, we modeled this "free-lunch learning" effect using artificial neural networks, in which the synaptic strength between each pair of model neurons is a connection weight. We proved that if forgetting is induced by allowing each weight value to drift randomly, then free-lunch learning is almost inevitable. However, if, after learning a set of associations, forgetting is induced by allowing each connection weight to *decay* or *fall* toward zero, then relearning a subset of associations decreases performance on the remaining associations. This suggests that evolution may have selected physiological mechanisms that involve forgetting using a form of synaptic drift rather than synaptic decay, because synaptic drift yields free-lunch learning, whereas decay does not.

"fall" towards the origin, then relearning some associations disrupts other associations.

The protocol used to examine FLL here is the same as that used in [4] and [12] and 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 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 error 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.

In order to preclude a common misunderstanding, we emphasize that, for a network with n connection weights, it is assumed that $n \geq n_1+n_2$; that is, the number of connection weights on each output unit is not less than the number n_1+n_2 of learned associations. Using the class of linear network models described below, up to n associations can be learned perfectly (see [12]).

The proofs below refer to a network with one output unit. However, these proofs apply to networks with multiple output units, because the n 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.

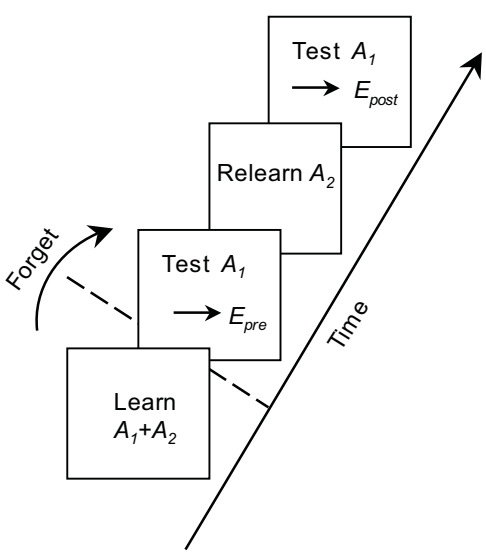

Figure 1. Free-lunch learning protocol. Two subsets of associations A_1 and A_2 are learned. After partial forgetting (see text), performance error E_{pre} on subset A_1 is measured. Subset A_2 is then relearned to pre-forgetting levels of performance, and performance error E_{post} on subset A_1 is re-measured. If $E_{post}<E_{pre}$ then FLL has occurred, and the amount of FLL is $\delta=E_{pre}-E_{post}$. Redrawn from [12].
doi:10.1371/journal.pcbi.1000143.g001

Definition of Performance Error

Each association consists of an input vector $\mathbf{x}$ and a corresponding target value d. For a network with weight vector $\mathbf{w}$, the response to an input vector $\mathbf{x}$ is $y=\mathbf{w}\cdot\mathbf{x}$. We define the *performance error* for input vectors $\mathbf{x}_1,\ldots,\mathbf{x}_k$ and desired outputs $d_1,\ldots,d_k$ to be

$$E(\mathbf{x}_1,\ldots,\mathbf{x}_k;\mathbf{w},d_1,\ldots,d_k)=\sum_{i=1}^{k}(y_i-d_i)^2, \tag{1}$$

where $y_i=\mathbf{w}\cdot\mathbf{x}_i$ is the output response to the input vector $\mathbf{x}_i$. By putting $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_k)^T$, $\mathbf{d}=(d_1,\ldots,d_k)^T$ and

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

we can write Equation 1 succinctly as

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

The two subsets A_1 and A_2 consist of n_1 and n_2 associations, respectively. Let $\mathbf{w}_0$ be the network weight vector after A_1 and A_2 are learned. When A_1 and A_2 are forgotten, the network weight vector changes to $\mathbf{w}_1$, say, and the performance error on A_1 becomes $E_{pre}=E(\mathbf{X};\mathbf{w}_1,\mathbf{d})$. Finally, relearning A_2 yields a new weight vector, $\mathbf{w}_2$, say, and the performance error on A_1 is $E_{post}=E(\mathbf{X};\mathbf{w}_2,\mathbf{d})$. Free-lunch learning has occurred if performance error on A_1 is less after relearning A_2 than it was before relearning A_2 (i.e., if $E_{post}<E_{pre}$).

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

$$\delta(\mathbf{w}_1,\mathbf{w}_2;\mathbf{X},\mathbf{d})=E_{pre}-E_{post} \tag{3}$$

which we shall also refer to simply as δ.

In previous work [12], we assumed that the "forgetting vector" $\mathbf{v}$ (defined as $\mathbf{v}=\mathbf{w}_1-\mathbf{w}_0$) has an isotropic distribution. Here we shall assume instead that the post-forgetting weight vector $\mathbf{w}_1$ is given by

$$\mathbf{w}_1=r\mathbf{w}_0 \tag{4}$$

for some (possibly random) scalar r, so that

$$\mathbf{v}=(r-1)\mathbf{w}_0 \tag{5}$$

and therefore

$$\mathbf{w}_1=\mathbf{w}_0-(1-r)\mathbf{w}_0 \tag{6}$$

The interpretation of Equation 6 is that forgetting consists of making the optimal weight vector $\mathbf{w}_0$ "fall" towards the origin by a *falling factor* $1-r$.

Results

We provide theoretical results, and compare these with results obtained using computer simulations. In essence, our theoretical and simulation results indicate that falling weights induce negative FLL, which decreases with the square of the falling factor $1-r$.

Theoretical Results

Our two main theorems are summarised here, and proofs are provided in the Methods section. These theorems apply to a network with n weights which learns n_1+n_2 associations $A=A_1\cup A_2$, and then after partial forgetting, relearns the n_2 associations in A_2.

We prove that if $n_1+n_2\leq n$ (so that, in general, the associations A_1 and A_2 are consistent) and the joint distribution of $(\mathbf{X}_1,\mathbf{d}_1)$ is isotropic (where $\mathbf{X}_1$ and $\mathbf{d}_1$ are the matrix of inputs and the vector of desired outputs for subset A_1 of associations) then the expected value of δ is negative (recall that δ is defined in Equation 3). We then prove that the probability $P(\delta<0)$ that δ is negative approaches unity as n_1 approaches ∞.

Theorem 1

For every non-zero value of r, the expected value of δ given r is negative. More precisely,

$$\mathbf{E}[\delta|r]\propto-(1-r)^2\frac{n_1}{n}, \tag{7}$$

with equality only in trivial cases, and where the constant of proportionality is guaranteed to be positive. Thus, the expected amount of FLL is negative (or zero).

From a physiological perspective, the case $r<1$ is obviously of interest because it represents synaptic weight decay. However, from a mathematical perspective, Theorem 1 applies to every value of r, and so it also holds for $r>1$. In other words, any movement of the weight vector $\mathbf{w}$ along the the line connecting $\mathbf{w}_0$ to the origin yields an expectation of negative FLL, in accordance with Theorem 1.

Theorem 2

Under mild conditions on the distributions of the input/output pairs $(\mathbf{X}_1,\mathbf{d}_1)$ and $(\mathbf{X}_2,\mathbf{d}_2)$,

$$P(\delta(\mathbf{w}_1,\mathbf{w}_2;\mathbf{X}_1,\mathbf{d}_1)\geq 0)\leq\frac{n}{n_1^2}\left(\frac{4\mathbf{E}\left[\|\tilde{\mathbf{x}}\|^2\right]}{\mathbf{E}\left[\|\mathbf{x}\|^2\right]}\mathbf{E}\left[\|\mathbf{d}_1\|^2\right]\mathbf{E}\left[\|\mathbf{d}_2\|^{-2}\right]+\frac{n_1[2(n-1)+3n\gamma(n)]}{n(n+2)}\right), \tag{8}$$

where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are any columns of $\mathbf{X}_1^T$ and $\mathbf{X}_2^T$, respectively, and

$$\gamma(n)=\frac{\mathbf{var}\left(\|\mathbf{x}\|^2\right)}{\mathbf{E}\left[\|\mathbf{x}\|^2\right]^2}.$$

Theorem 2 implies that, if (i) the number (n_1) of associations in A_1 is a fixed non-zero proportion (n_1/n) of the number n of connection weights, (ii) $\mathbf{E}[\|\mathbf{d}_1\|^2]\mathbf{E}[\|\mathbf{d}_2\|^{-2}]$ is bounded as $n\rightarrow\infty$, and (iii) $\gamma(n)\rightarrow 0$ as $n\rightarrow\infty$ then $P(\delta>0)\rightarrow 0$ as $n\rightarrow\infty$, i.e., the amount of FLL is negative, with a probability which tends to 1 as $n\rightarrow\infty$.

For example, if we assume that (i) each input vector $\mathbf{x}=(x_1,\ldots,x_n)$ is chosen from an isotropic Gaussian distribution and (ii) the variance of x_i is σ_x^2 then $\gamma(n)=2/n$, $\mathbf{E}\left[\|\mathbf{x}\|^2\right]=\mathbf{E}\left[\|\tilde{\mathbf{x}}\|^2\right]$, and $\mathbf{E}[\|\mathbf{d}_1\|^2]\mathbf{E}[\|\mathbf{d}^2\|^{-2}]=n_1/(n_2-1)$. This ensures that $P(\delta>0)\rightarrow 0$ as $n\rightarrow\infty$.

Simulation Results

Simulation was carried out on 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 comprised 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 component of the ith input vector $\mathbf{x}_i$, and each weight w_j is the connection between the jth input unit and the output unit.

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 $\mathbf{w}_{\text{new}}=\mathbf{w}_{\text{old}}-\eta\nabla E_{(\mathbf{w}_{\text{old}})}$, where η is the learning rate, which is adjusted according to the number of weights.

However, in order to save time, we used an equivalent learning method. 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 the weight vector $\mathbf{w}_0$ was obtained; this provided perfect performance on all n associations. Partial forgetting was induced by making weights "fall" towards the origin $\mathbf{w}_1=r\mathbf{w}_0$, after which performance error was E_{pre}. Relearning the $n_2=n/2$ associations in A_2 was implemented with $k=n_2$ as above, after which performance error was E_{post}.

In each simulation, each value in each input vector $\mathbf{x}_i$, and each target value d_i was chosen from the same isotropic gaussian distribution with unit variance. There were 100 input units, and one output unit. The subsets A_1 and A_2 each consisted of 50 associations. The value of $\delta=E_{\text{pre}}-E_{\text{post}}$ was obtained in each of 100 simulations, using a different random seed for each simulation. In Figure 2, the mean of 100 values of δ is shown for various values of the falling factor $1-r$.

The Geometry of Forgetting

We present a brief account of the geometry which underpins the results reported here, for a network with two input units and one output unit, as shown in Figure 3A. This network learns two associations $A_1=(X_1,d_1)$ and $A_2=(X_2,d_2)$.

Figure 3B provides a geometric example of how relearning A_2 increases the error on A_1. After learning A_1 and A_2, $\mathbf{w}=\mathbf{w}_0$. The effects of forgetting and relearning can be seen by ignoring the $\pm$ superscripts and subscripts for now. After partial forgetting, $\mathbf{w}=\mathbf{w}_1$, and performance error $E_{\text{pre}}=p^2$. Relearning A_2 yields $\mathbf{w}_2$, the orthogonal projection of $\mathbf{w}_1$ on to L_2, and performance error is $E_{\text{post}}=q^2$. FLL occurs if $\delta=E_{\text{pre}}-E_{\text{post}}>0$, or equivalently if $p^2-q^2>0$ (see [12], Appendices A–C for proofs). Forgetting here consists of reducing $\mathbf{w}_0$ by a factor $r<1$, so that $\mathbf{w}_1=r\mathbf{w}_0$.

The plus and minus signs in Figure 3B refer to two versions A_1^+ and A_1^- of association A_1, in which X_1 is the same and the target d_1 has the same magnitude, but opposite signs: $A_1^+=(X_1,+d_1)$ and $A_1^-=(X_1,-d_1)$.

We now find the expected change in error induced by relearning a given association A_2. After learning $\{A_1^+,A_2\}$ followed by forgetting, the change in error on A_1^+ after relearning A_2 is $\delta^+=\delta(\mathbf{w}_1^+,\mathbf{w}_2^+;X_1,+d_1)$. After learning $\{A_1^-,A_2\}$ followed by forgetting, the change in error on A_1^- after relearning A_2 is $\delta^-=\delta(\mathbf{w}_1^-,\mathbf{w}_2^-;X_1,-d_1)$. Using similar triangles in Figure 3B,

$$p_+=(1-r)d_1,\ q_+=(1-r)(d_1-e) \tag{9}$$

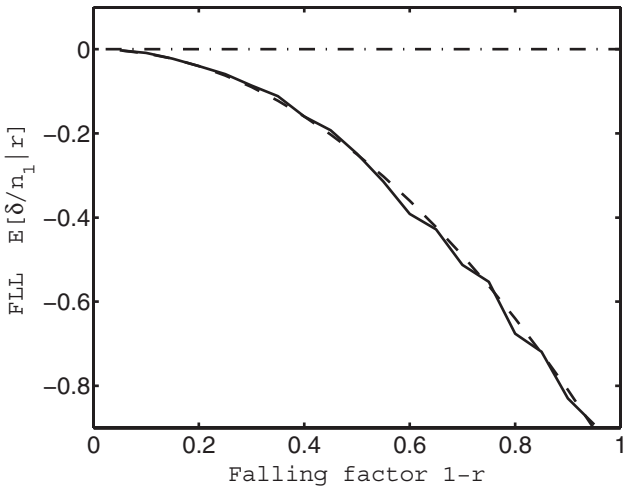

Figure 2. Free-lunch learning decreases as the network's weight vector falls toward the origin. A network with 100 input units and one output unit learns two subsets A_1 and B_2, each of which consists of 50 associations. After learning A_1 and A_2, the network has a weight vector $\mathbf{w}=\mathbf{w}_0$, but after partial forgetting, the weight vector is $\mathbf{w}=\mathbf{w}_1$. If forgetting consists of subtracting a proportion $1-r$ of $\mathbf{w}_0$ such that $\mathbf{w}_1=\mathbf{w}_0-(1-r)\mathbf{w}_0$ then the weight vector "falls" towards the origin; the factor $1-r$ is called the *falling factor*. After forgetting, performance error on A_1 is E_{pre}, an error which changes to E_{post} after relearning A_2, where this change is $\delta=E_{\text{pre}}-E_{\text{post}}$. Given that there are A_1 associations in A_1, the expected free-lunch learning per association in A_1 is therefore $E[\delta/n_1|r]$. *Solid curve*: the expected FLL, $E[\delta/n_1|r]$, where this expectation is taken over 100 computer simulations. *Dashed curve*: theoretical prediction of $E[\delta/n_1|r]$ (see Equation 7), using a constant of proportionality equal to unity, so that the predicted free-lunch learning is $E_{\text{predict}}[\delta/n_1|r]=-(1-r)^2$. As predicted, free-lunch learning $E[\delta/n_1|r]$ becomes more negative as the falling factor $1-r$ increases.
doi:10.1371/journal.pcbi.1000143.g002

$$p_-=(1-r)d_1,\ q_-=(1-r)(d_1+e) \tag{10}$$

Therefore, the total change in error on A_1^+ and A_1^- induced by relearning A_2 (on different occasions) is

$$\delta^+ + \delta^- = \left(p_+^2 - q_+^2\right) + \left(p_-^2 - q_-^2\right) \tag{11}$$

$$= -2(1-r)^2 e^2 \tag{12}$$

$$<0 \tag{13}$$

Irrespective of the precise value of the target output value d_1 in A_1, if the distribution of d_1 is isotropic then $+d_1$ is as probable as $-d_1$. If the total change in error for two instances (A_1^+ and A_1^-) of A_1 is $-2(1-r)^2e^2$ then the expected change (conditional on e) is $E[\delta\,|\,e]=-(1-r)^2e^2$. Therefore, if forgetting is induced by falling weight values, then the expected change in error $\mathbf{E}[\delta]<0$.

Discussion

We have proved and demonstrated that, in one of the simplest forms of neural network model, relearning part of a previously learned set of associations reduces performance on the remaining non-relearned associations. This result is in stark contrast to our previous results, which proved that relearning induced partial recovery of non-relearned items [12]. The only difference between these two studies is the way in which forgetting was induced.

An obvious physiological concomitant of Hebbian learning is long-term potentiation (LTP), which seems to underpin learned behaviors [14]. LTP can last for hours, days or even months, and usually follows an exponential decay [3]. However, some forms of LTP do not seem to decay [15], and have been shown to be stable for up to one year [16]. Such stability is remarkable, but from a statistical point of view, would almost certainly be accompanied by random fluctuations which would have a cumulative effect over time; and indeed, fluctuations are apparent in the stable LTP reported in [16]. Crucially, it is not known if the forgetting of learned behaviors is caused by decaying efficacy at many synapses, or by the cumulative effect of random fluctuations in stable LTP-induced synaptic efficacies. Here, decaying efficacy is analogous to weight values that fall toward zero in network models, whereas the cumulative effect of random fluctuations is analogous to the addition of random noise, or drifting, of weight values in network models.

Given a choice between forgetting via synaptic weights that fall towards zero and weights that drift isotropically, has evolution chosen drifting or falling? If all other things were equal then forgetting via synaptic drift would seem to be the obvious choice. This is because drifting ensures that relearning a subset of associations improves performance on other associations, whereas falling decreases performance. However, other things are rarely equal. The expected magnitude of weights increases with drifting but decreases with falling. (Consider a hypersphere centered on the origin, with radius $\|\mathbf{w}_0\|$. Simple geometry shows that more than half of all directions emanating from $\mathbf{w}_0$ yield a new weight vector $\mathbf{w}_1$ which lies outside the hypersphere, and therefore $\mathbf{E}[\|\mathbf{w}_1\|]>\mathbf{E}[\|\mathbf{w}_0\|]$ (assuming, for example, that all vectors $\mathbf{w}_1-\mathbf{w}_0$ have the same length).) This decrease in weight magnitudes effectively reduces neuronal firing rates, which reduces metabolic costs relative to costs incurred by synaptic drift. Synaptic drift therefore confers mnemonic benefits, but these benefits come at a metabolic price. Thus the increased fitness gained from the mnemonic benefits of synaptic drift must be offset against their metabolic costs. In essence, even free-lunch learning comes at a price.

Methods

We proceed by deriving expressions for E_{pre}, E_{post}, and for $\delta=E\text{pre}-E_{\text{post}}$. We prove that if $n_1+n_2\leq n$ then the expected value of δ is negative. We then prove that the probability $P(\delta<0)$ that δ is negative approaches unity as n_1 approaches ∞.

Performance Errors

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

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

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

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

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

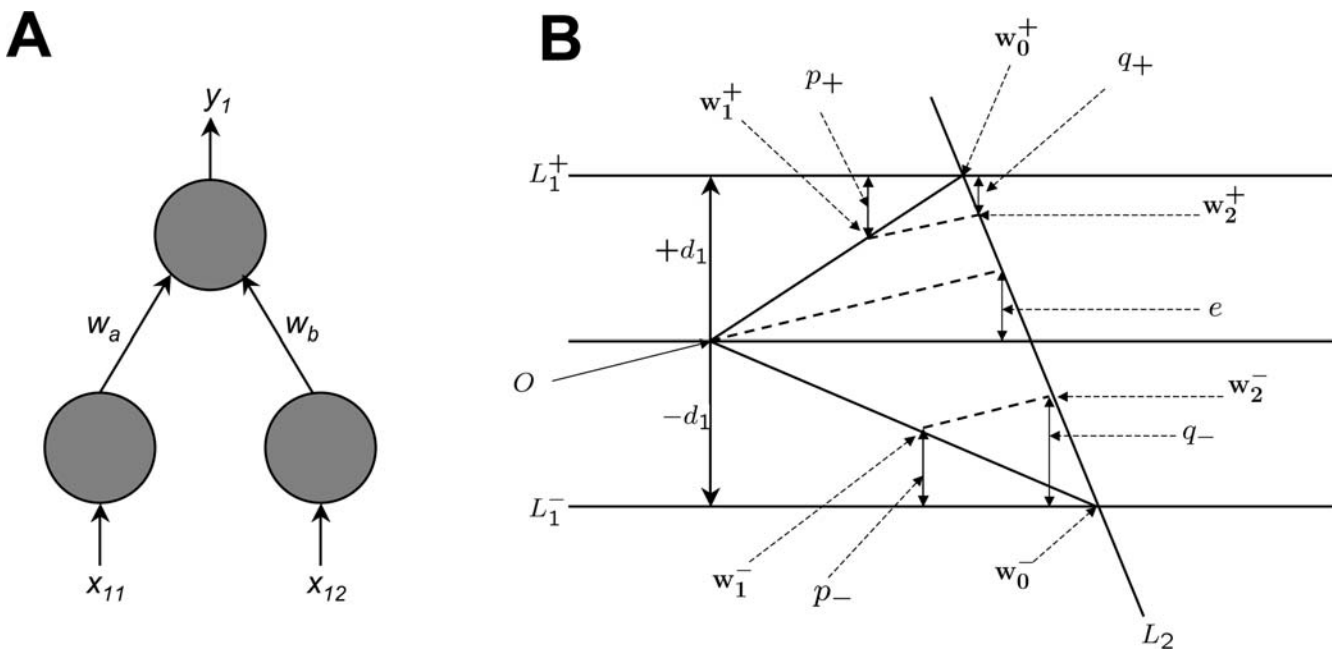

Figure 3. Geometric example of how relearning A_2 increases the error on A_1. (A) A network with two input units and one output unit, with connection weights ω_a and ω_b defines a weight vector $\mathbf{w}=(\omega_a,\omega_b)$. The network learns two associations A_1 and A_2. For example, A_1 is the mapping from input vector $\mathbf{x}_1=(x_{11},x_{12})$ to desired output value d_1, and learning A_1 consists of adjusting $\mathbf{w}$ until the network output $y_1=\mathbf{w}\cdot\mathbf{x}_1$ equals d_1. (B) For a given association $A_2=(X_2,d_2)$, the corresponding constraint line in the space defined by (ω_a,ω_b) is L_2. Irrespective of the precise value of the target output value d_1 in association A_1, if d_1 is distributed isotropically then $+d_1$ is as probable as $-d_1$. When averaged over $+d_1$ and $-d_1$, the change δ in error on A_1 induced by relearning A_2 can be shown to be $-(1-r)^2e^2$, where $\mathbf{w}_1^{\pm}=r\mathbf{w}_0^{\pm}$. Since this is less than zero, the expected change $E[\delta|r]<0$. (Figure 3A redrawn from [12]).
doi:10.1371/journal.pcbi.1000143.g003

of $L_{\mathbf{X},\mathbf{d}}$. Then

$$\begin{aligned}\delta(\mathbf{w}_1,\mathbf{w}_2;\mathbf{X},\mathbf{d}) &= \|\mathbf{X}\mathbf{w}_1-\mathbf{d}\|^2-\|\mathbf{X}\mathbf{w}_2-\mathbf{d}\|^2\\ &= \|\mathbf{X}\mathbf{w}_1\|^2-\|\mathbf{X}\mathbf{w}_2\|^2-2(\mathbf{w}_1-\mathbf{w}_2)^T\mathbf{X}^T\mathbf{d}\\ &= (\mathbf{w}_1-\mathbf{w}_2)^T\mathbf{X}^T\mathbf{X}(\mathbf{w}_1+\mathbf{w}_2)-2(\mathbf{w}_1-\mathbf{w}_2)^T\mathbf{X}^T\mathbf{X}\tilde{\mathbf{w}}\\ &= (\mathbf{w}_1-\mathbf{w}_2)^T\mathbf{X}^T\mathbf{X}(\mathbf{w}_1+\mathbf{w}_2-2\tilde{\mathbf{w}}).\end{aligned} \tag{14}$$

If $\mathbf{X}_i$ has rank n_i then transposing the QR decomposition of $\mathbf{X}_i^T$ (or, equivalently, using Gram–Schmidt orthonormalisation of the rows of $\mathbf{X}_i$) gives

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

for unique $n_i\times n_i$ and $n_i\times n$ matrices $\mathbf{T}_i$ and $\mathbf{Z}_i$ with $\mathbf{T}_i$ lower triangular with positive diagonal elements, and $\mathbf{Z}_i\mathbf{Z}_i^T=\mathbf{I}_{n_i}$. Simple calculation shows that, for any weight vector $\mathbf{w}$, $\left(\mathbf{I}_n-\mathbf{Z}_i\mathbf{Z}_i^T\right)\mathbf{w}$ and $\mathbf{Z}_i\mathbf{Z}_i^T\mathbf{w}$ are orthogonal. Since $\mathbf{w}=\left(\mathbf{I}_n-\mathbf{Z}_i\mathbf{Z}_i^T\right)\mathbf{w}+\mathbf{Z}_i\mathbf{Z}_i^T\mathbf{w}$, it follows that the matrix $\mathbf{Z}_i^T\mathbf{Z}_i$ represents the operator that projects orthogonally onto the image of $\mathbf{Z}_i^T\mathbf{Z}_i$. Because

$$\mathbf{Z}_i^T\mathbf{Z}_i\mathbf{X}_i^T\mathbf{X}_i=\mathbf{X}_i^T\mathbf{X}_i, \tag{15}$$

the image of $\mathbf{X}_i^T\mathbf{X}_i$ is contained in that of $\mathbf{Z}_i^T\mathbf{Z}_i$. As both these images have dimension n_i, they must be equal, and so $\mathbf{Z}_i^T\mathbf{Z}_i$ represents the operator which projects orthogonally onto the image of $\mathbf{X}_i^T\mathbf{X}_i$.

Now suppose that $\mathbf{X}$ and $\mathbf{d}$ are consistent, where

$$\mathbf{X}=\begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{pmatrix}\quad \mathbf{d}=\begin{pmatrix}\mathbf{d}_1\\ \mathbf{d}_2\end{pmatrix}.$$

Then, after the network has learned A_1 and A_2, the weight vector $\mathbf{w}_0$ satisfies

$$\mathbf{X}_1\mathbf{w}_0=\mathbf{d}_1 \text{ and } \mathbf{X}_2\mathbf{w}_0=\mathbf{d}_2 \tag{16}$$

(If, as below, $n_1+n_2\leq n$, $\mathbf{X}_2$ and $\mathbf{d}_2$ are consistent, and $(\mathbf{X}_1,\mathbf{d}_1)$ has a continuous distribution then Equation 16 holds with probability 1.)

Falling

We now assume that forgetting is induced by weight values "falling" towards the origin at zero, i.e., forgetting consists of shrinking the weight vector $\mathbf{w}_0$ by a (possibly random) factor r towards the "dead state" $\mathbf{0}$. Thus the post-forgetting weight vector $\mathbf{w}_1$ is given by

$$\mathbf{w}_1=r\mathbf{w}_0 \tag{17}$$

and so the "forgetting vector" $\mathbf{v}=\mathbf{w}_1-\mathbf{w}_0$ is

$$\mathbf{v}=(r-1)\mathbf{w}_0 \tag{18}$$

The form of forgetting given by Equation 17 is very different from that investigated in [12], where $\mathbf{v}$ has an isotropic distribution and is independent of $(\mathbf{X}_1,\mathbf{d}_1)$ and $(\mathbf{X}_2,\mathbf{d}_2)$.

Let $\mathbf{w}_2$ be the orthogonal projection of $\mathbf{w}_1$ onto L_2. Then

$$\mathbf{w}_2 = \mathbf{w}_0 + \left(\mathbf{I}_n - \mathbf{Z}_2^T\mathbf{Z}_2\right)(\mathbf{w}_1 - \mathbf{w}_0).$$

Manipulation gives

$$\mathbf{w}_1 - \mathbf{w}_2 = \mathbf{Z}_2^T\mathbf{Z}_2\mathbf{v}, \tag{19}$$

and so

$$\mathbf{w}_1 + \mathbf{w}_2 - 2\mathbf{w}_0 = \left(2\mathbf{I}_n - \mathbf{Z}_2^T\mathbf{Z}_2\right)\mathbf{v}. \tag{20}$$

Then Equations 14, 16, and 18–20 yield

$$\begin{aligned}
&\delta(\mathbf{w}_1,\mathbf{w}_2;\mathbf{X}_1,\mathbf{d}_1)\\
&= \mathbf{v}^T\mathbf{Z}_2^T\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\left(2\mathbf{I}_n - \mathbf{Z}_2^T\mathbf{Z}_2\right)\mathbf{v}\\
&= (1-r)^2\left(\mathbf{T}_2^{-1}\mathbf{d}_2\right)^T\mathbf{Z}_2\mathbf{X}_1^T\left(2\mathbf{d}_1 - \mathbf{X}_1\mathbf{Z}_2^T\mathbf{T}_2^{-1}\mathbf{d}_2\right)\\
&= (1-r)^2\left\{2\left(\mathbf{T}_2^{-1}\mathbf{d}_2\right)^T\mathbf{Z}_2\mathbf{X}_1^T\mathbf{d}_1 - \left(\mathbf{T}_2^{-1}\mathbf{d}_2\right)^T\mathbf{Z}_2\mathbf{X}_1^T\mathbf{X}_1\mathbf{Z}_2^T\mathbf{T}_2^{-1}\mathbf{d}_2\right\}
\end{aligned} \tag{21}$$

The Case of Isotropic Random $(\mathbf{X}_1,\mathbf{d}_1)$

In this section we assume that the distribution of $(\mathbf{X}_1,\mathbf{d}_1)$ is isotropic, i.e., that $(\mathbf{U}\mathbf{X}_1\mathbf{V},\mathbf{U}\mathbf{d}_1)$ has the same distribution as $(\mathbf{X}_1,\mathbf{d}_1)$ for all orthogonal $n_1\times n_1$ matrices $\mathbf{U}$ and all orthogonal $n\times n$ matrices $\mathbf{V}$. Then taking the conditional expectation of Equation 21 for given $\mathbf{X}_2$, $\mathbf{d}_2$, and r gives the following theorem.

Theorem 1

If

1. $n_1+n_2\leq n$,
2. $\mathbf{X}_2$ and $\mathbf{d}_2$ are consistent,
3. the distribution of $(\mathbf{X}_1,\mathbf{d}_1)$ is continuous and isotropic,
4. $\mathbf{X}_1$, $\mathbf{d}_1$, and $(\mathbf{X}_2,\mathbf{d}_2,r)$ are independent.

then

$$\mathbf{E}[\delta(\mathbf{w}_1,\mathbf{w}_2;\mathbf{X}_1,\mathbf{d}_1)|\mathbf{X}_2,\mathbf{d}_2,r] = -(1-r)^2\frac{n_1}{n}\mathbf{E}\left[\|\mathbf{x}\|^2\right]\left\|\mathbf{T}_2^{-1}\mathbf{d}_2\right\|^2, \tag{22}$$

where $\mathbf{x}$ is any column of $\mathbf{X}_1^T$.

Corollary 1

If 1.-3. of Theorem 1 hold then

$$\mathbf{E}[\delta(\mathbf{w}_1,\mathbf{w}_2;\mathbf{X}_1,\mathbf{d}_1)|\mathbf{X}_2,\mathbf{d}_2,r] \leq 0 \tag{23}$$

with equality if and only if either $r=1$ or $\mathbf{d}_2=0$.

Corollary 1 says that (apart from trivial exceptions) the expected amount of FLL is negative.

To obtain Theorem 2, it is useful to have some moments of isotropic distributions. Let $\mathbf{x}$ be isotropically distributed on $\mathbb{R}^n$. Then Equations 9.6.1 and 9.6.2 of Mardia and Jupp (2000), together with some algebraic manipulation, yield

$$\mathbf{E}\left[\mathbf{x}^T\mathbf{A}\mathbf{x}\right] = \frac{\mathbf{E}\left[\|\mathbf{x}\|^2\right]\ \mathbf{tr}(\mathbf{A})}{n} \tag{24}$$

$$\begin{aligned}
\mathbf{var}\left(\mathbf{x}^T\mathbf{A}\mathbf{x}\right) &= \frac{\mathbf{E}\left[\|\mathbf{x}\|^4\right]\ \left\{n\mathbf{tr}\left(\mathbf{A}^2\right) + n\mathbf{tr}\left(\mathbf{A}\mathbf{A}^T\right) - 2\mathbf{tr}(\mathbf{A})^2\right\}}{n^2(n+2)}\\
&\quad + \frac{\mathbf{var}\left(\|\mathbf{x}\|^2\right)\ \mathbf{tr}(\mathbf{A})^2}{n^2},
\end{aligned} \tag{25}$$

as in Equations A.14 and A.15 of [12].

The other tool used in proving Theorem 2 is the formula

$$\mathrm{var}(Y|X) = \mathbf{E}[\mathrm{var}(Y|X,Z)|Z] + \mathrm{var}(\mathbf{E}[Y|X,Z]|Z) \tag{26}$$

for any random variables X,Y,Z for which these quantities exist. Equation 26 is an application to the conditional distribution of $Y|Z$ of the standard conditional variance formula that is given in Equation 2b.3.6 on page 97 of [17].

Taking the expectation and variance of Equation 21 as only $\mathbf{d}_1$ varies and using Equation 24 gives

$$\begin{aligned}
&\mathbf{E}[\delta(\mathbf{w}_1,\mathbf{w}_2;\mathbf{X}_1,\mathbf{d}_1)|\mathbf{X}_1,\mathbf{X}_2,\mathbf{d}_2,r]\\
&= -(1-r)^2\left(\mathbf{Z}_2^T\mathbf{T}_2^{-1}\mathbf{d}_2\right)^T\mathbf{X}_1^T\mathbf{X}_1\left(\mathbf{Z}_2^T\mathbf{T}_2^{-1}\mathbf{d}_2\right),
\end{aligned} \tag{27}$$

$$\begin{aligned}
&\mathrm{var}(\delta(\mathbf{w}_1,\mathbf{w}_2;\mathbf{X}_1,\mathbf{d}_1)|\mathbf{X}_1,\mathbf{X}_2,\mathbf{d}_2,\mathbf{r})\\
&= 4(1-r)^4\frac{\mathbf{E}\left[\|\mathbf{d}_1\|^2\right]}{n_1}\left(\mathbf{Z}_2^T\mathbf{T}_2^{-1}\mathbf{d}_2\right)^T\mathbf{X}_1^T\mathbf{X}_1\left(\mathbf{Z}_2^T\mathbf{T}_2^{-1}\mathbf{d}_2\right).
\end{aligned} \tag{28}$$

Taking the expectation of Equation 28 as only $\mathbf{X}_1$ varies and using Equation 24 gives

$$\begin{aligned}
&\mathbf{E}[\mathbf{var}(\delta(\mathbf{w}_1,\mathbf{w}_2;\mathbf{X}_1,\mathbf{d}_1)|\mathbf{X}_1,\mathbf{X}_2,\mathbf{d}_2,r)|\mathbf{X}_2,\mathbf{d}_2,r]\\
&= 4(1-r)^4\frac{\mathbf{E}\left[\|\mathbf{d}_1\|^2\right]\mathbf{E}\left[\|\mathbf{x}\|^2\right]}{n}\left\|\mathbf{T}_2^{-1}\mathbf{d}_2\right\|^2.
\end{aligned} \tag{29}$$

We now suppose that

$$\text{the columns } \mathbf{x}_1,\ldots,\mathbf{x}_{n1} \text{ of } \mathbf{X}_1^T \text{ are distributed independently.} \tag{30}$$

Then taking the variance of Equation 27 as only $\mathbf{X}_1$ varies and using Equation 25 gives

$$\begin{aligned}
&\mathbf{var}(\mathbf{E}[\delta(\mathbf{w}_1,\ \mathbf{w}_2;\ \mathbf{X}_1,\ \mathbf{d}_1)|\mathbf{X}_1,\ \mathbf{X}_2,\ \mathbf{d}_2,\ r]\ |\mathbf{X}_2,\ \mathbf{d}_2,\ r)\\
&= n_1(1-r)^4\frac{\left\|\mathbf{T}_2^{-1}\mathbf{d}_2\right\|^4}{n^2}\left\{\mathbf{E}\left[\|\mathbf{x}\|^4\right]\ \frac{2(n-1)}{n+2} + \mathbf{var}\left(\|\mathbf{x}\|^2\right)\right\}.
\end{aligned} \tag{31}$$

Adding Equations 29 and 30 and using Equation 26 yields

$$\begin{aligned}
&\mathbf{var}(\delta(\mathbf{w}_1,\mathbf{w}_2;\mathbf{X}_1,\mathbf{d}_1)|\mathbf{X}_2,\mathbf{d}_2,r)\\
&= (1-r)^4\frac{\left\|\mathbf{T}_2^{-1}\mathbf{d}_2\right\|^2}{n}\times\left(4\mathbf{E}\left[\|\mathbf{d}_1\|^2\right]\mathbf{E}\left[\|\mathbf{x}\|^2\right] + \frac{n_1}{n}\left\|\mathbf{T}_2^{-1}\mathbf{d}_2\right\|^2\right.\\
&\left.\left\{\mathbf{E}\left[\|\mathbf{x}\|^4\right]\frac{2(n-1)}{n+2} + \mathbf{var}\left(\|\mathbf{x}\|^2\right)\right\}\right).
\end{aligned} \tag{32}$$

To obtain an upper bound on the conditional probability of FLL (i.e., on $P(\delta \geq 0 \mid \mathbf{X}_2, \mathbf{d}_2, r)$), we use Chebyshev's inequality, which states that, for any random variable Y and any positive value of t

$$P(|Y - \mathbf{E}[Y]| \geq t) \leq \frac{\mathbf{var}(Y)}{t^2}.$$

Applying Chebyshev's inequality to the conditional distribution of $\delta(\mathbf{w}_1, \mathbf{w}_2, \mathbf{X}_1, \mathbf{d}_1)$ given $(\mathbf{X}_2, \mathbf{d}_2, r)$, taking $t = \mathbf{E}[\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \mid \mathbf{X}_2, \mathbf{d}_2, r]$, and noting that (by Equation 23) $t \leq 0$, we obtain

$$P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \geq 0 \mid \mathbf{X}_2, \mathbf{d}_2, r) \leq \frac{\mathbf{var}(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \mid \mathbf{X}_2, \mathbf{d}_2, r)}{\mathbf{E}[\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \mid \mathbf{X}_2, \mathbf{d}_2, r]^2}. \tag{33}$$

Substituting Equations 22 and 32 into Equation 33 gives

$$P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \geq 0 \mid \mathbf{X}_2, \mathbf{d}_2, r) \leq \frac{n}{n_1^2} \left(\frac{4\mathbf{E}\left[\|\mathbf{d}_1\|^2\right]}{\|\mathbf{T}_2^{-1}\mathbf{d}_2\|^2 \mathbf{E}\left[\|\mathbf{x}\|^2\right]} + \frac{n_1[2(n-1) + 3n\gamma(n)]}{n(n+2)} \right), \tag{34}$$

where

$$\gamma(n) = \frac{\mathbf{var}\left(\|\mathbf{x}\|^2\right)}{\mathbf{E}\left[\|\mathbf{x}\|^2\right]^2}.$$

For any positive-definite symmetric matrix $\mathbf{A}$ and vector $\mathbf{x}$, diagonalization of $\mathbf{A}$, together with the fact that $x + 1/x \geq 2$ for positive x, yields

$$\left(\mathbf{x}^T \mathbf{A} \mathbf{x}\right)\left(\mathbf{x}^T \mathbf{A}^{-1} \mathbf{x}\right) \geq \|\mathbf{x}\|^4 \tag{35}$$

Combining Equations 34 and 35 with the fact that $\mathbf{T}_2 \mathbf{T}_2^T = \mathbf{X}_2 \mathbf{X}_2^T$ gives

$$P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \geq 0 \mid \mathbf{X}_2, \mathbf{d}_2, r) \leq \frac{n}{n_1^2} \left(\frac{4\mathbf{E}\left[\|\mathbf{d}_1\|^2\right] \mathbf{d}_2^T \mathbf{X}_2 \mathbf{X}_2^T \mathbf{d}_2}{\|\mathbf{d}_2\|^4 \mathbf{E}\left[\|\mathbf{x}\|^2\right]} + \frac{n_1[2(n-1) + 3n\gamma(n)]}{n(n+2)} \right), \tag{36}$$

Taking the expectation of Equation 36 over $\mathbf{X}_2$ yields

$$P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \geq 0 \mid \mathbf{d}_2, r) \leq \frac{n}{n_1^2} \left(\frac{4\mathbf{E}\left[\|\mathbf{d}_1\|^2\right] \mathbf{E}\left[\|\tilde{\mathbf{x}}\|^2\right]}{\|\mathbf{d}_2\|^2 \mathbf{E}\left[\|\mathbf{x}\|^2\right]} + \frac{n_1[2(n-1) + 3n\gamma(n)]}{n(n+2)} \right), \tag{37}$$

where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are any columns of $\mathbf{X}_1^T$ and $\mathbf{X}_2^T$, respectively.

Taking the expectation of Equation 37 over $\mathbf{d}_2$ and r yields the following theorem.

Theorem 2

If (a) conditions 1.-4. of Theorem 1 hold, (b) the columns $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}$ of $\mathbf{X}_1^T$ are distributed independently, (c) $\mathbf{X}_2$, $\mathbf{d}_2$, and r are independent, (d) the distribution of $(\mathbf{X}_2, \mathbf{d}_2)$ is isotropic, and (e) $\mathbf{E}[\|\mathbf{d}_2\|^{-2}]$ is finite then

$$P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \geq 0) \leq \frac{n}{n_1^2} \left(\frac{4\mathbf{E}\left[\|\tilde{\mathbf{x}}\|^2\right]}{\mathbf{E}\left[\|\mathbf{x}\|^2\right]} \mathbf{E}\left[\|\mathbf{d}_1\|^2\right] \mathbf{E}\left[\|\mathbf{d}_2\|^{-2}\right] + \frac{n_1[2(n-1) + 3n\gamma(n)]}{n(n+2)} \right), \tag{38}$$

where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are any columns of $\mathbf{X}_1^T$ and $\mathbf{X}_2^T$, respectively, and

$$\gamma(n) = \frac{\mathbf{var}\left(\|\mathbf{x}\|^2\right)}{\mathbf{E}\left[\|\mathbf{x}\|^2\right]^2}.$$

Corollary 2

If the conditions of Theorem 2 hold and

$$\mathbf{x} \sim N\left(\mathbf{0}, \sigma_x^2 \mathbf{I}_n\right),\ \mathbf{d}_1 \sim N\left(\mathbf{0}, \sigma_x^2 \mathbf{I}_{n_1}\right),$$
$$\tilde{\mathbf{x}} \sim N\left(\mathbf{0}, \sigma_x^2 \mathbf{I}_n\right),\ \mathbf{d}_2 \sim N\left(\mathbf{0}, \sigma_x^2 \mathbf{I}_{n_2}\right),$$

where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are any columns of $\mathbf{X}_1^T$ and $\mathbf{X}_2^T$, respectively, then

$$P(\delta(\mathbf{w}_1, \mathbf{w}_2; \mathbf{X}_1, \mathbf{d}_1) \geq 0) \leq \frac{2(2n + n_2 - 2)}{n_1(n_2 - 2)}.$$

Thus

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

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

Acknowledgements

Thanks to David Sterratt for asking, "What would happen to free-lunch learning if weights decayed?" and to three anonymous reviewers for their detailed comments.

Author Contributions

Conceived and designed the experiments: JS. Performed the experiments: JS. Analyzed the data: JS. Contributed reagents/materials/analysis tools: JS. Wrote the paper: JS PEJ. Mathematical proofs: PEJ.

References

1. Tanzi E (1893) I fatti e le induzioni nellodierna isologia del sistema nervosa. Riv Sper Freniatr Med Leg 19: 419–472.
2. Hebb D (1949) The Organization of Behavior: A Neuropsychological Theory. New York: Wiley.

3. Abraham W (2003) How long will long-term potentiation last? Philos Trans R Soc Lond B Biol Sci 358: 735–744.
4. Stone J, Hunkin N, Hornby A (2001) Predicting spontaneous recovery of memory. Nature 414: 167–168.
5. Coltheart M, Byng S (1989) A treatment for surface dyslexia. In: Seron X, ed. Cognitive Approaches in Neuropsychological Rehabilitation. London: Lawrence Erlbaum Associates.
6. Weekes B, Coltheart M (1996) Surface dyslexia and surface dysgraphia: treatment studies and their theoretical implications. Cogn Neuropsychol 13: 277–315.
7. Atkins P (2001) What happens when we relearn part of what we previously knew? Predictions and constraints for models of long-term memory. Psychol Res 65: 202–215.
8. Roediger H III (1973) Inhibition in recall from cueing with recall targets. J Verbal Learn Verbal Behav 12: 644–657.
9. Serra M, Nairne J (2000) Part-set cuing of order information: implications for associative theories. Mem Cognit 28: 847–855.
10. Hinton G, Plaut D (1987) Using fast weights to deblur old memories. In: Proceedings Ninth Annual Conference of the Cognitive Science Society. pp 177–186.
11. Atkins P, Murre J (1998) Recovery of unrehearsed items in connectionist models. Connect Sci 10: 99–119.
12. Stone J, Jupp P (2007) Free-lunch learning: modelling spontaneous recovery of memory. Neural Comput 19: 194–217.
13. Stone J (2007) Distributed representations accelerate evolution of adaptive behaviours. PLoS Comput Biol 3: e147. doi:10.1371/journal.pcbi.0030147.
14. Whitlock J, Heynen A, Shuler M, Bear M (2006) Learning induces long-term potentiation in the hippocampus. Science 313: 1093–1097.
15. Staubli U, Lynch G (1987) Stable hippocampal long-term potentiation elicited by theta pattern stimulation. Brain Res 435: 227–234.
16. Abraham WC, Logan B, Greenwood JM, Dragunow M (2002) Induction and experience-dependent consolidation of stable long-term potentiation lasting months in the hippocampus. J Neurosci 22: 9626–9634.
17. Rao C (1973) Linear Statistical Inference and its Applications. 2nd edition. New York: Wiley.