Cognition 230 (2023) 105290

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Discussion

Hebbian, correlational learning provides a memory-less mechanism for Statistical Learning irrespective of implementational choices: Reply to Tovar and Westermann (2022)☆

Ansgar D. Endress [a,*], Scott P. Johnson [b]

[a] Department of Psychology, City, University of London, UK
[b] Department of Psychology, University of California, Los Angeles, United States of America

ARTICLE INFO

Keywords:
Statistical learning
Implicit learning
Transitional probabilities
Neural networks
Chunking

ABSTRACT

Statistical learning relies on detecting the frequency of co-occurrences of items and has been proposed to be crucial for a variety of learning problems, notably to learn and memorize words from fluent speech. Endress and Johnson (2021) (hereafter EJ) recently showed that such results can be explained based on simple memory-less correlational learning mechanisms such as Hebbian Learning. Tovar and Westermann (2022) (hereafter TW) reproduced these results with a different Hebbian model. We show that the main differences between the models are whether temporal decay acts on both the connection weights and the activations (in TW) or only on the activations (in EJ), and whether interference affects weights (in TW) or activations (in EJ). Given that weights and activations are linked through the Hebbian learning rule, the networks behave similarly. However, in contrast to TW, we do not believe that neurophysiological data are relevant to adjudicate between abstract psychological models with little biological detail. Taken together, both models show that different memory-less correlational learning mechanisms provide a parsimonious account of Statistical Learning results. They are consistent with evidence that Statistical Learning might not allow learners to learn and retain words, and Statistical Learning might support predictive processing instead.

Statistical learning relies on detecting the frequency of co-occurrences of items, and has been proposed to be crucial for a variety of learning problems (e.g. Aslin, Saffran, & Newport, 1998; Kirkham, Slemmer, & Johnson, 2002; Morgan, Fogel, Nair, & Patel, 2019; Saffran, Aslin, & Newport, 1996; Saffran & Griepentrog, 2001; Saffran, Newport, & Aslin, 1996; Stalinski & Schellenberg, 2010; Turk-Browne & Scholl, 2009; Verosky & Morgan, 2021), notably learning words from fluent speech (Aslin & Newport, 2012; Aslin et al., 1998; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). We recently showed that such results can be explained based on simple correlational learning mechanisms such as Hebbian Learning (Endress & Johnson, 2021) (hereafter EJ). Tovar and Westermann (2022) (hereafter TW) reproduced these results with a slightly different model (with temporal decay acting on both the connection weights and the activations, rather than on only the activations, and interference affecting weights rather than activations), and offering different interpretations of some network parameters (e.g., conceiving of forgetting as decay).

Here, we first stress the common theoretical implications of both models: While Statistical Learning is often assumed to help learners learn (and thus *memorize*) words from fluent speech (e.g. Erickson, Thiessen, & Estes, 2014; Graf-Estes, Evans, Alibali, & Saffran, 2007; Isbilen, McCauley, Kidd, & Christiansen, 2020; Karaman & Hay, 2018; Shoaib, Wang, Hay, & Lany, 2018), results from the tasks typically

☆ The code used in this article is available at https://github.com/aendress/tp_model_reply_to_tw and https://doi.org/10.25383/city.20054993. This research was supported by NIH, United States of America grant R01-HD073535 to SPJ.

* Corresponding author.
E-mail address: ansgar.endress.1@city.ac.uk (A.D. Endress).

[1] Historically, many authors have stressed the importance of correlational learning mechanisms (if not exactly Hebb's rule), from Hume's 1739/2003 theory of causation to collocation detection in natural language processing (Manning & Schütze, 1999), though other authors questioned whether what appears to be correlational learning (e.g., conditioning) really reflects correlational learning *mechanisms* (Gallistel & Gibbon, 2000). Our main assumption is that learning mechanisms that show properties of correlational learning (e.g., the effects of Hebbian learning) are *psychologically* plausible even though there are many other methods of detecting co-occurrences.

https://doi.org/10.1016/j.cognition.2022.105290
Received 12 June 2022; Received in revised form 30 August 2022; Accepted 17 September 2022
Available online 11 October 2022

used to explore Statistical Learning can be explained by a memory-less correlational learning model. As a result, Statistical Learning might be more useful for predictive processing than for learning words *per se* (e.g. Endress & de Seyssel; Morgan et al., 2019; Sherman & Turk-Browne, 2020; Turk-Browne, Scholl, Johnson, & Chun, 2010; Verosky & Morgan, 2021). Following this, we briefly discus the differences between EJ's and TW's models. As already argued by EJ, we agree that different implementations of correlational learning are likely to result in fairly similar results.[1] However, we also show that, contrary to TW's characterization of their model, activation decay is critical to their model's performance, and argue that models of psychological phenomena should be evaluated by their psychological predictions rather than by reference to their putative "biological plausibility" when neither model includes biophysical attributes.

1. A memory-less interpretation of statistical learning

One of the primary motivations of Statistical Learning is that it might allow learners to extract (and memorize) words from fluent speech (e.g. Aslin & Newport, 2012; Aslin et al., 1998; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). Speech is often thought to be a continuous signal (but see Brentari, González, Seidl, & Wilbur, 2011; Christophe, Mehler, & Sebastian-Galles, 2001; Endress & Hauser, 2010; Johnson & Jusczyk, 2001; Johnson & Seidl, 2009; Pilon, 1981; Shukla, Nespor, & Mehler, 2007; Shukla, White, & Aslin, 2011). As a result, to acquire any word, learners first need to learn where words start and where they end. To this end, they might use Transitional Probabilities (TPs) among syllables, that is, the conditional probability of a syllable σ_{i+1} given a preceding syllable σ_i, $P(\sigma_i\sigma_{i+1})/P(\sigma_i)$. Unpredictable transitions might indicate a word boundary, while relatively predictable transitions are likely located inside words. Humans are sensitive to TPs (Aslin et al., 1998; Kirkham et al., 2002; Morgan et al., 2019; Saffran, Aslin, & Newport, 1996; Saffran & Griepentrog, 2001; Saffran, Newport, & Aslin, 1996; Stalinski & Schellenberg, 2010; Turk-Browne & Scholl, 2009), and might use this sensitivity to memorize words (e.g. Erickson et al., 2014; Graf-Estes et al., 2007; Isbilen et al., 2020; Karaman & Hay, 2018; Shoaib et al., 2018).

However, the evidence that Statistical Learning leads to memory for words is mixed at best (see Endress, Slone, & Johnson, 2020 for a critical review). For example, when exposed to statistically structured sequences, participants are sometimes more familiar with high-TP items than with low-TP items, even when they have never encountered either of them and thus could not have memorized them (because the items are played backwards with respect to the familiarization sequence; Endress & Wood, 2011; Jones & Pashler, 2007; Turk-Browne & Scholl, 2009). In other cases, participants are more familiar with high-TP items they have *never* heard or seen than with low-TP items they have encountered (Endress & Langus, 2017; Endress & Mehler, 2009). Further, when instructed to repeat back the items they remember from a statistically structured familiarization sequences, participants are unable to do so even when they learned the statistical structure of the stream (Endress & de Seyssel).

Such results thus suggest that Statistical Learning abilities do not necessarily support the formation of declarative memories for words. This interpretation mirrors earlier demonstrations of dissociations between Statistical Learning and declarative memory (e.g. Cohen & Squire, 1980; Finn et al., 2016; Graf & Mandler, 1984; Poldrack et al., 2001; Squire, 1992), and suggests that Statistical Learning might be more useful for predictive processing rather than declarative memory formation (e.g. Endress & de Seyssel; Morgan et al., 2019; Sherman & Turk-Browne, 2020; Turk-Browne et al., 2010; Verosky & Morgan, 2021).[2] To the extent that Statistical Learning has a computational function (in Marr and Nishihara's 1992 sense, and is not a spandrel, Gould, Lewontin, Maynard Smith, & Holliday, 1979), we thus surmise that its function is the prediction of future events.

Both EJ's and TW's models are consistent with this view. EJ simulated the results of a number of Statistical Learning results with a fully connected network where the strength of excitatory connections among neurons was tuned by Hebbian learning. That is, if two neurons are active simultaneously, their connection becomes strengthened ("what fires together wires together"). The network also comprised inhibitory connections among neurons. Further, the network had a "forgetting" mechanism, where activity decayed as time passed. After familiarization with a speech stream, the network was tested by recording the total activation when presented with different types of test items.

The basic result was that this fairly generic network accounted for a number of Statistical Learning results. Critically, given that all learning resided in the connection strengths, it could do so without any memory representations at all. In fact, just as in human participants (Endress & Langus, 2017; Endress & Mehler, 2009), the network activation was determined by the associative strength of the syllables in an item, irrespective of whether the network had encountered the item or not. As a result, the network had no memory representation of either item (or one would need to conclude that the network remembered items it has never encountered).

EJ also found that, to account for these Statistical Learning results, the forgetting rate needed to be reasonable. Rather unsurprisingly, if forgetting was so fast that neurons were never active together, no learning ensued. Conversely, if forgetting was so slow that all neurons were active simultaneously, all neurons formed connections, making these indiscriminate connections useless as an indicator of learning.

2. Differences between EJ's and TW's model

TW reproduced these results in a similar network, confirming that basic Hebbian learning mechanisms can explain Statistical Learning results, to some extent independently of how they are implemented. As far as we can see, there are four main differences between TW's and EJ's models. First, TW take issue with our characterization of decay as forgetting. Second, TW stress the importance of spreading activation. Third, TW evaluate learning by inspecting connections rather than activations. Fourth, instead of including separate inhibitory and decay/forgetting components that affect activations (and thus indirectly connection weights through the Hebbian learning rule), their model uses a modified Hebbian learning rule (with an additional parameter) where decay/forgetting affects weights (and thus indirectly activations); this learning rule also comprises a thresholding mechanism that presumably mimics the effects of mutual inhibition.

2.1. Forgetting vs. decay

Regarding the interpretation of EJ's "forgetting" parameter, TW "argue that [interpreting decay as forgetting] may be a misleading interpretation. Activation values from external stimuli in both artificial and biological networks are non-persistent but are constantly updated in response to changes in the environment (Huber & O'Reilly, 2003)". We agree that, in our specific implementation, decay would also be a reasonable description of the phenomenon we tried to capture. That

[2] Under the premise that Statistical Learning is used for word learning, the finding that Statistical Learning might not require declarative memory might lead to the conclusion that word learning does not require declarative memory either. However, given the *prima facie* plausibility of the view that words need to be retrieved from memory (at least for production), it seems more plausible that declarative memory is involved in word learning, and that Statistical Learning abilities might thus not support the formation of memories for words. We thank an anonymous reviewer for suggesting this possibility.

said, given the controversy over whether decay plays a role in forgetting (see below), we believe that forgetting is a theoretically more neutral term, especially because the same phenomenon can likely be captured by manipulating the interference parameter rather than the decay parameter (see below). However, we would question to what extent results from single neuron recordings are relevant for *psychological* models that are not particularly plausible biologically to begin with; for example both EJ's and TW's "neurons" code for speaker-independent, phonological representations of syllables, which would presumably be encoded by some fairly abstract population code in actual brains (Pouget, Dayan, & Zemel, 2000). Further, while decay has certainly been widely documented, so has persistent neural activity, which exists in various brain areas and taxa (e.g. Major & Tank, 2004). As a result, neurophysiological findings may not be informative about psychological theories.

In fact, the question of whether time-based decay exists in memory is a controversial one in cognitive psychology. Under some circumstances, humans can remember thousands of items for hours or weeks (Brady, Konkle, Alvarez, & Oliva, 2008; Standing, 1973); under other circumstances, very similar pictures disappear from memory after a few seconds but can be reviewed through repeated exposure (Endress & Potter, 2014; Pertzov & Avidan, 2009; Thunell & Thorpe, 2019). Further, it is controversial whether there is any decay in Short-Term Memory at all, or whether all decreases in memory are due to interference (e.g. Baddeley & Scott, 1971; Berman, Jonides, & Lewis, 2009; Lewandowsky, Oberauer, & Brown, 2009; Nairne, Whiteman, & Kelley, 1999). We are thus open to different psychological interpretations of the forgetting parameter, and EJ already acknowledged the possibility that the effects of their forgetting parameter could likely be mimicked by tuning inhibition (see below).

In contrast, although TW argue that their "simulation results … challenge E&J's notion of activation decay as the key ingredient for Hebbian statistical learning", forgetting/decay is critical to their model. They use decay in two places. First, the activation of each input is maintained only for two time steps (at 90% for the second time step); given that the current input is likely the strongest activation at each time step, the effects are similar to a global forgetting parameter. Second, TW consider only activation greater than a certain threshold. While the effect of the latter seems to be a reduced overall magnitude of the weights, the former is critical for the results. To illustrate this fact, we exposed the network to the familiarization stream from Saffran, Aslin, and Newport's 1996 Experiment 2, and then recorded the weights in high-TP items ("words") and low-TP items (part-words, of BC:D and C:DE type, a difference that is irrelevant for the current purposes). We ran 1000 simulations with three version of TW's model: With the original decay function from TW ("Standard" in Fig. 1), no forgetting at all (i.e., the input to each neurons was the cumulative sum of prior inputs; "Never" in Fig. 1) and immediate forgetting (i.e., the activation decays immediately after presentation; "Immediate" in Fig. 1). As shown in Fig. 1, the network discriminated between words and part-words only using TW's decay function; as in EJ's simulations, all weights reach the maximum of 1.0 in the absence of decay, and reached zero with immediate forgetting. A suitably chosen decay parameter is thus crucial to TW's model. Be that as it might, we believe that the merits of psychological models should be evaluated by their empirical adequacy, and links between psychological parameters and neurobiological findings should be investigated empirically.

2.2. The role of spreading activation

TW stress the importance for spreading activation for network performance. We certainly agree, and, in their Section 2, EJ explained the role of spreading activation in detail. In fact, we suggested that spreading activation might be a more parsimonious account of previous claims that Statistical Learning might lead to word-like memory representations (see Erickson et al., 2014; Graf-Estes et al., 2007; Isbilen et al., 2020; Karaman & Hay, 2018; Shoaib et al., 2018 vs. Endress & Langus, 2017; Endress et al., 2020).

Given the importance of spreading activation, it is surprising that TW evaluate their model by inspecting connections weights rather than by measuring activations. In fact, even in a network with uniform connections and no learning, it is hard to describe the network dynamics mathematically without resorting to simulation (Endress & Szabó, 2020). Given that, in TW's model, interference and decay act on weights rather than activations, this problem might be somewhat reduced in their model. Still, just relying on the pattern of weights, it is hard to evaluate the dynamic interplay of first and higher order associations or the dynamic aspects of statical learning such as those revealed by electrophysiological recordings (Endress & Fló, in preparation).

2.3. The effects of different learning rules

2.3.1. One vs. two component learning rules

The most critical difference between EJ and TWs models is the learning rule. TW's learning rule has two components. First, all weights undergo decay. This decay is proportional to the current weight and the product of the activations connected by that weight, that is

$$\Delta_{\text{Decay}} W_{AB} \propto -W_{AB} \times \text{activation}_A \times \text{activation}_B,$$

where A and B are two neurons. However, given that, even in the simple Hebbian learning rule

$$\Delta W_{AB} \propto \text{activation}_A \times \text{activation}_B$$

the weight change is proportional to product of the activations, the effects of decay on learning will be very similar irrespective of whether decay originates from weights, activations or, as in TW's model, both. However, in the absence of targeted experiments investigating the empirical adequacy of weight-based vs activation-based decay, the key result is that both formalisms account for Statistical Learning results in the absence of a memory mechanism.

The second component of TW's learning rule is the strengthening of associations according to the simple Hebbian learning rule above. Critically, however, TW's model strengthens connections only when the product of the activation exceeds an arbitrary threshold ($\text{activation}_A \times \text{activation}_B > \theta$). However, the effect of this thresholding is similar to inhibitory connections. To see why this is the case, consider two pairs of neurons. The activations in each pair are roughly similar to each other, but the activation in the first pair is somewhat larger than in the second pair (i.e, $\text{activation}_A \approx \text{activation}_B > \text{activation}_C \approx \text{activation}_D$). If there is inhibition, the first pair will reduce the activation of the second pair as long as the inhibitory input exceeds their excitatory input (though the difference does not necessarily disappear; Endress & Szabó, 2020). Given that weight changes are proportional to the product of the corresponding activations, connections between neurons with greater coactivation will be strengthened to a greater extent, irrespective of whether this is implemented through inhibitory connections or through an co-activation-based threshold for learning.

Again, we believe that targeted psychological experiments are necessary to gauge the empirical adequacy of activation-based decay and inhibition (as in EJ) vs. weight-based decay and inhibition (as in TW). In fact, there is evidence for both kinds of processes. On the one hand, the type of lateral, activation-based inhibition assumed in EJ has been proposed as a psychological mechanism for phenomena from perception to attention to response inhibition (e.g. Desimone & Duncan, 1995; Hampshire & Sharp, 2015). On the other hand, to the extent that knowledge of TPs resides in connection weights, the finding that knowledge of TPs is forgotten after a few minutes (e.g. Karaman & Hay, 2018; Vlach & DeBrock, 2019) suggests that "weights" can be forgotten over time, which is consistent with TW's proposal of weight-based decay (though this forgetting might still occur through interference or decay).

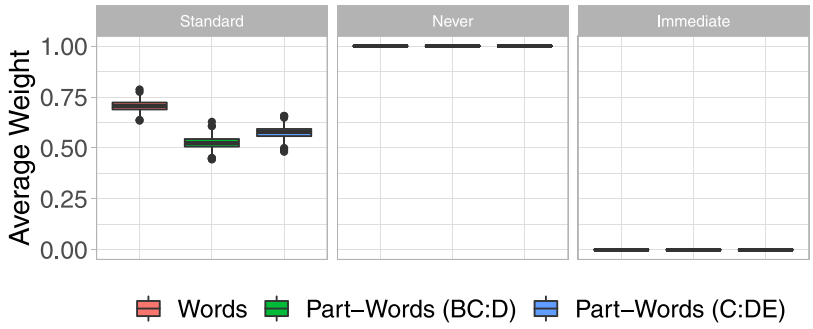

Fig. 1. Average connection weights of the test items in a simulation of Saffran, Aslin, and Newport's 1996 Experiment 2, using TW's model. (Left) Simulations using decay parameters from TW's model. (Middle) Simulations with no activation decay. (Right) Simulations with immediate decay. High-TP items (words) are discriminated from low-TP items (part-words of different types) only with a suitable decay function. With no decay, all weights are maximal; with immediate forgetting, no connections are formed.

For the current purposes, we just assume that some normalization mechanism keeps activations at a reasonable level, and believe that the question of whether normalization occurs through weight decay, weight interference or lateral inhibition is best answered through experimental rather than computational investigations. To the extent that biological plausibility is relevant for psychological models, the ubiquity of lateral inhibition across domains and taxa certainly suggest that activation-based inhibition is no less plausible than weight-based inhibition.

2.3.2. Connection weights do not grow excessively in the absence of weight forgetting

TW justified their two component learning rule in part by arguing that "it is not clear their [EJ's] model prevents excessive growth of connections". However, it is easy to see from EJ's Hebbian learning rule that the final weight of the connection between two neurons after t time steps is proportional to the average coactivation of the neurons, $W_{AB}(t) \propto t \times \langle \text{activation}_A \times \text{activation}_B \rangle$ (for $W_{AB}(0) = 0$). As a result, if the activations remain in a reasonable range, so will the weights. This is confirmed when examining the connection weights after familiarization with a stream modeled after Saffran, Aslin, and Newport's 1996 Experiment 2. As shown in Fig. 2, connection weights diverge for slow decay rates of up to .2, but generally stay below or around 1 for faster decay rates. In other words, weights stay in a reasonable range for decay rates that led to learning in EJ's simulations; for decay rates that were too slow for learning to occur, weights diverge as well. This confirms our point above that qualitatively similar results can be achieved by controlling weights (and thus indirectly activations, as in TW's simulations) or by controlling activations (and thus indirectly weights, as in EJ's simulations).

2.4. Decay vs. interference

TW questioned EJ's rationale for not varying their interference parameter. However, and as mentioned above, EJ argued that their "interference parameter might well mimic the role of forgetting", and thus simply sought to limit the number of moving parts in their model. To see why this is the case, consider a network of N neurons that receive external stimulation in a regular sequence. In the absence noise, the activation change between times t and $t+1$ is given by (exponential) decay (first term), spreading activation (second term), inhibition (third term) and external stimulation (fourth term).

$$x_i(t+1) - x_i(t) = -\lambda_a x_i(t) + \alpha \sum_{j \neq i} w_{ij} F(x_j) - \beta \sum_{j \neq i} F(x_j) + I(t) \quad (1)$$

To see the relationship between decay and inhibition, we assume that excitatory connectivity is relatively sparse, and partition the neurons into a set of K neurons with excitatory connections with target neuron i, and $N-K$ neurons with negligible excitatory input to neuron i.

$$\begin{aligned} x_i(t+1) - x_i(t) = &-\lambda_a x_i(t) + \alpha \sum_{j=1, j \neq i}^{K} w_{ij} F(x_j) \\ &- \beta \sum_{j=1, j \neq i}^{K} F(x_j) - \beta \sum_{j=K+1, j \neq i}^{N} F(x_j) \\ &+ I(t) \end{aligned} \quad (2)$$

In the absence of external stimulation, the $N - K$ neurons not providing excitatory input to i will provide periodic inhibitory input (Endress & Fló, in preparation) that is generally unrelated to the activation of i. Averaged across time, this input thus mimics the effect of linear (rather than exponential) decay. For the K neurons that provide excitatory input to i, their excitatory input is proportional to their inhibitory input. As the activation of i is also an increasing function of this excitatory input, the inhibition might thus mimic exponential decay (though the specific functional form is more complex). Further, given that i is presumably most active when closely associated neurons are active as well (and assuming sparse activations), this exponential-like inhibition is likely the dominant inhibitory input when i has noteworthy activation.

In the presence of external stimulation, the excitatory input to i from other neurons is no longer related to i's activation. However, given the symmetry of the Hebbian learning rule, i will also excite the neurons it has excitatory connections with. As a result, these neurons will again provide inhibitory input that is an increasing function of i's activation, albeit with a time-lag. Consequently, the effects of inhibition and time-based decay can likely mimic one another. Critically, however, given that EJ's objective was to make the conceptual point that Statistical Learning results can be reproduced by a simple, memory-less correlational learning mechanism, they did not explore alternative implementations of this idea. Be that as it may, TW's model confirms that EJ's results can be reproduced with different implementations.

3. Conclusions

In sum, both EJ and TW show that a memory-less correlational learning mechanism can account for results from Statistical Learning studies, despite differences in implementation, irrespective of whether decay and inhibition affect activations or weights.[3] As a result, to the extent that Statistical Learning supports declarative memory formation for words, relevant evidence is still required.

[3] A stronger argument for the implementation independence of our conclusions would rely on analytic results for classes of activation functions and learning rules. However, given that mathematical treatments of neural network properties (e.g., using statistical mechanics) usually assume *some* learning rule (e.g. Amit, 1989; Hopfield, 1982; Huang, 2021; Storkey, 1997), and that an analytic derivation of the asymptotic network behavior is challenging even for a simplified version of the current model with less complex stimuli (Endress

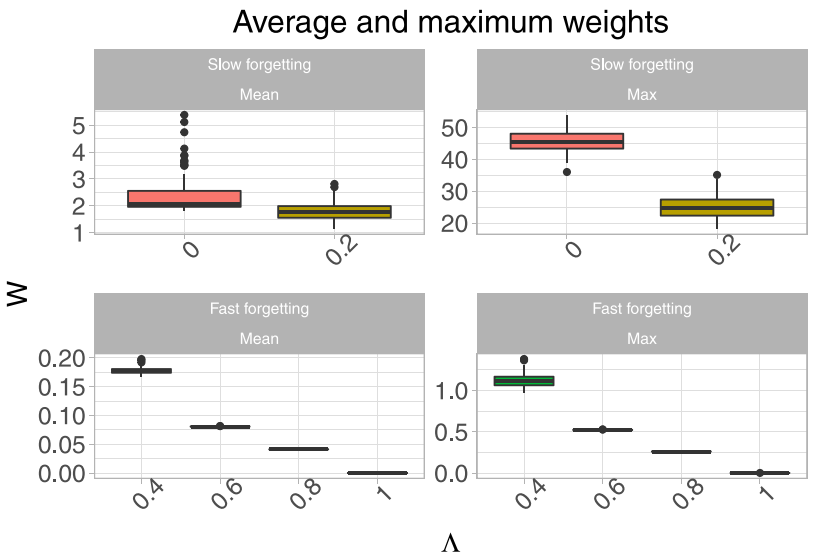

Fig. 2. Final weights after simulation of stream from Saffran, Aslin, and Newport's 1996, using EJ's model. Mean (left) and maximal (right) weights for slow (top) and fast (bottom) forgetting rates.

Data availability

Link to code is shared in manuscript

References

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

Aslin, R. N., & Newport, E. L. (2012). Statistical learning. *Current Directions in Psychological Science*, *21*(3), 170–176. http://dx.doi.org/10.1177/0963721412436806.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Baddeley, A. D., & Scott, D. (1971). Short term forgetting in absence of proactive interference. *The Quarterly Journal of Experimental Psychology*, *23*, 275–283.

Berman, M. G., Jonides, J., & Lewis, R. L. (2009). In search of decay in verbal short-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(2), 317–333. http://dx.doi.org/10.1037/a0014873.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14325–14329. http://dx.doi.org/10.1073/pnas.0803390105.

Brentari, D., González, C., Seidl, A., & Wilbur, R. (2011). Sensitivity to visual prosodic cues in signers and nonsigners. *Language and Speech*, *54*(1), 49–72.

Christophe, A., Mehler, J., & Sebastian-Galles, N. (2001). Perception of prosodic boundary correlates by newborn infants. *Infancy*, *2*(3), 385–394.

Cohen, N., & Squire, L. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*, *210*(4466), 207–210. http://dx.doi.org/10.1126/science.7414331.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222. http://dx.doi.org/10.1146/annurev.ne.18.030195.001205.

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177–199. http://dx.doi.org/10.1016/j.cogpsych.2010.05.001.

Endress, A. D., & Johnson, S. P. (2021). When forgetting fosters learning: A neural network model for statistical learning. *Cognition*, Article 104621. http://dx.doi.org/10.1016/j.cognition.2021.104621.

Endress, A. D., & Langus, A. (2017). Transitional probabilities count more than frequency, but might not be used for memorization. *Cognitive Psychology*, *92*, 37–64. http://dx.doi.org/10.1016/j.cogpsych.2016.11.004.

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, *60*(3), 351–367. http://dx.doi.org/10.1016/j.jml.2008.10.003.

Endress, A. D., & Potter, M. C. (2014). Something from (almost) nothing: Buildup of object memory from forgettable single fixations. *Attention, Perception and Psychophysics*, *76*(8), 2413–2423. http://dx.doi.org/10.3758/s13414-014-0706-3.

& Szabó, 2020), we believe that testing different implementations is important to strengthen the generality of our argument.

Endress, A D., & de Seyssel, M. under review. The specificity of sequential Statistical Learning: Statistical Learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory. http://dx.doi.org/10.31234/osf.io/u9z4a.

Endress, A. D., Slone, L. K., & Johnson, S. P. (2020). Statistical learning and memory. *Cognition*, *204*, Article 104346. http://dx.doi.org/10.1016/j.cognition.2020.104346.

Endress, A. D., & Szabó, S. (2020). Sequential presentation protects memory from catastrophic interference. *Cognitive Science*, *44*(5), http://dx.doi.org/10.1111/cogs.12828.

Endress, A. D., & Wood, J. N. (2011). From movements to actions: Two mechanisms for learning action sequences. *Cognitive Psychology*, *63*(3), 141–171. http://dx.doi.org/10.1016/j.cogpsych.2011.07.001.

Erickson, L. C., Thiessen, E. D., & Estes, K. G. (2014). Statistically coherent labels facilitate categorization in 8-month-olds. *Journal of Memory and Language*, *72*, 49–58. http://dx.doi.org/10.1016/j.jml.2014.01.002.

Finn, A. S., Kalra, P. B., Goetz, C., Leonard, J. A., Sheridan, M. A., & Gabrieli, J. D. (2016). Developmental dissociation between the maturation of procedural memory and declarative memory. *Journal of Experimental Child Psychology*, *142*, 212–220. http://dx.doi.org/10.1016/j.jecp.2015.09.027.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, *107*(2), 289–344.

Gould, S. J., Lewontin, R. C., Maynard Smith, J., & Holliday, R. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *205*(1161), 581–598. http://dx.doi.org/10.1098/rspb.1979.0086.

Graf, P., & Mandler, G. (1984). Activation makes words more accessible, but not necessarily more retrievable. *Journal of Verbal Learning and Verbal Behavior*, *23*(5), 553–568. http://dx.doi.org/10.1016/s0022-5371(84)90346-3.

Graf-Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, *18*(3), 254–260. http://dx.doi.org/10.1111/j.1467-9280.2007.01885.x.

Hampshire, A., & Sharp, D. J. (2015). Contrasting network and modular perspectives on inhibitory control. *Trends in Cognitive Sciences*, *19*, 445–452. http://dx.doi.org/10.1016/j.tics.2015.06.006.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*, 2554–2558. http://dx.doi.org/10.1073/pnas.79.8.2554.

Huang, H. (2021). *Statistical mechanics of neural networks*. Singapore: Springer Nature, http://dx.doi.org/10.1007/978-981-16-7570-6.

Huber, D. E., & O'Reilly, R. C. (2003). Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. *Cognitive Science*, *27*(3), 403–430. http://dx.doi.org/10.1207/s15516709cog2703_4.

Hume, D. (1739/2003). *A treatise of human nature*. Project Gutenberg.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science*, *44*, Article e12848. http://dx.doi.org/10.1111/cogs.12848.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567.

Johnson, E. K., & Seidl, A. H. (2009). At 11 months, prosody still outranks statistics. *Developmental Science*, *12*(1), 131–141. http://dx.doi.org/10.1111/j.1467-7687.2008.00740.x.

Jones, J., & Pashler, H. (2007). Is the mind inherently forward looking? Comparing prediction and retrodiction. *Psychonomic Bulletin & Review*, *14*, 295–300. http://dx.doi.org/10.3758/bf03194067.

Karaman, F., & Hay, J. F. (2018). The longevity of statistical learning: When infant memory decays, isolated words come to the rescue. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(2), 221–232. http://dx.doi.org/10.1037/xlm0000448.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. http://dx.doi.org/10.1016/s0010-0277(02)00004-5.

Lewandowsky, S., Oberauer, K., & Brown, G. D. A. (2009). No temporal decay in verbal short-term memory. *Trends in Cognitive Sciences*, *13*(3), 120–126. http://dx.doi.org/10.1016/j.tics.2008.12.003.

Major, G., & Tank, D. (2004). Persistent neural activity: Prevalence and mechanisms. *Current Opinion in Neurobiology*, *14*(6), 675–684. http://dx.doi.org/10.1016/j.conb.2004.10.017.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marr, D., & Nishihara, H. K. (1992). Visual information processing: Artificial intelligence and the sensorium of sightrtificial intelligence and the sensorium of sight. In S. M. Kosslyn, & R. A. Andersen (Eds.), *Frontiers in cognitive neuroscience* (pp. 165–186). Cambridge, MA: MIT Press, Reprinted from Technology Review 81:2–23 (1978).

Morgan, E., Fogel, A., Nair, A., & Patel, A. D. (2019). Statistical learning and gestalt-like principles predict melodic expectations. *Cognition*, *189*, 23–34. http://dx.doi.org/10.1016/j.cognition.2018.12.015.

Nairne, J. S., Whiteman, H. L., & Kelley, M. R. (1999). Short-term forgetting of order under conditions of reduced interference. *The Quarterly Journal of Experimental Psychology*, *52*, 241–251.

Pertzov, Y., & Avidan, E. (2009). Accumulation of visual information across multiple fixations. *Journal of Vision*, *9*(10), 2.1–212. http://dx.doi.org/10.1167/9.10.2.

Pilon, R. (1981). Segmentation of speech in a foreign language. *Journal of Psycholinguistic Research*, *10*(2), 113–122.

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., et al. (2001). Interactive memory systems in the human brain. *Nature*, *414*, 546–550. http://dx.doi.org/10.1038/35107080.

Pouget, A., Dayan, P., & Zemel, R. S. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, *1*(2), 125–132.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology*, *37*(1), 74–85.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cueshe role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Sherman, B. E., & Turk-Browne, N. B. (2020). Statistical prediction of the future impairs episodic encoding of the present. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 22760–22770. http://dx.doi.org/10.1073/pnas.2013291117.

Shoaib, A., Wang, T., Hay, J. F., & Lany, J. (2018). Do infants learn words from statistics? Evidence from English-learning infants hearing Italian. *Cognitive Science*, *42*(8), 3083–3099. http://dx.doi.org/10.1111/cogs.12673.

Shukla, M., Nespor, M., & Mehler, J. (2007). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*(1), 1–32. http://dx.doi.org/10.1016/j.cogpsych.2006.04.002.

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(15), 6038–6043. http://dx.doi.org/10.1073/pnas.1017617108.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*(2), 195–231. http://dx.doi.org/10.1037/0033-295x.99.2.195.

Stalinski, S. M., & Schellenberg, E. G. (2010). Shifting perceptions: Developmental changes in judgments of melodic similarity. *Developmental Psychology*, *46*(6), 1799–1803. http://dx.doi.org/10.1037/a0020658.

Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology*, *25*(2), 207–222.

Storkey, A. (1997). Increasing the capacity of a hopfield network without sacrificing functionality. In W. Gerstner, A. Germond, M. Hasler, & J. D. Nicoud (Eds.), *Artificial neural networks* (pp. 451–456). Berlin, Heidelberg: Springer Berlin Heidelberg.

Thunell, E., & Thorpe, S. J. (2019). Memory for repeated images in rapid-serial-visual-presentation streams of thousands of images. *Psychological Science*, *30*, 989–1000. http://dx.doi.org/10.1177/0956797619842251.

Tovar, A. E., & Westermann, G. (2022). No need to forget, just keep the balance: Hebbian neural networks for statistical learning. *Cognition*.

Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(1), 195–202.

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *Journal of Neuroscience*, *30*, 11177–11187. http://dx.doi.org/10.1523/JNEUROSCI.0858-10.2010.

Verosky, N. J., & Morgan, E. (2021). Pitches that wire together fire together: Scale degree associations across time predict melodic expectations. *Cognitive Science*, *45*, Article e13037. http://dx.doi.org/10.1111/cogs.13037.

Vlach, H. A., & DeBrock, C. A. (2019). Statistics learned are statistics forgotten: Children's retention and retrieval of cross-situational word learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*(4), 700–711.