Hebbian, correlational learning provides a memory-less mechanism for Statistical Learning irrespective of implementational choices: Reply to Tovar & Westermann

Ansgar D. Endress
Scott P. Johnson
Department of Psychology, City, University of London, UK
Department of Psychology, UCLA

Draft of June 9, 2022

Abstract

TO BE WRITTEN

Keywords: Statistical Learning; Implicit Learning; Transitional Probabilities; Neural Networks; Chunking

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., ?, ?, ?, ?, ?, ?, ?, ?, ?, ?), notably learning words from fluent speech (?, ?, ?, ?, ?). We recently showed that such results can be explained based on simple correlational learning mechanisms such as Hebbian Learning (?, ?) (hereafter EJ). ? (?) (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., ?, ?, ?, ?, ?, ?, ?), the tasks 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*

An `R` implementation of the model is available at `XXX`. Please note that the URL will change in the final version of the manuscript. The final location will be `XXX`. This research was support by NIH grant R01-HD073535 to SPJ.

se (e.g., ?, ?, ?, ?, ?, ?). 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. 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 no biophysical detail whatsoever.

1 A memory-less interpretion 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., ?, ?, ?, ?, ?). Speech is often thought to be a continuous signal (but see ?, ?, ?, ?, ?, ?, ?, ?, ?). As a result, to acquire any word, learners first need to learn where words start and where they 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 (?, ?, ?, ?, ?, ?, ?, ?, ?), and might use this sensitivity to memorize words (e.g., ?, ?, ?, ?, ?, ?).

However, the evidence that Statistical Learning leads to memory for words is mixed at best (see ?, ? 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; ?, ?, ?, ?). 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 (?, ?, ?). 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 (?, ?).

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., ?, ?, ?, ?, ?, ?), and suggests that Statistical Learning might be more useful for predictive processing rather than declarative memory formation (e.g., ?, ?, ?, ?, ?, ?).

Both EJ's and TW's are consistent with this view. EJ simulated the 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 humans participants (?, ?, ?), the network activation was determined by the associative strength of the syllables in a 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 rememebered 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 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 mechanisms that presumably mimics the effects of mutual inhibition.

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 (?, ?)." While we are not particularly committed to the label "forgetting" and while persistent neural activity has been widely documented in various brain areas (?, ?), we would question to what extent results from single neuron recordings are relevant for *psychological* models that are not particularly plausible biologically; 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 (?, ?). 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 (?, ?, ?); under other circumstances, very simi-

Figure 1. Average connection weights of the test items in a simulation of ?'s (?) Experiment 2, using TW's model. (Left) Simulations using decay parameter's 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 differen types) only with a suitable decay function. With no decay, all weights are maximal; with immediate forgetting, no connections are formed.

lar pictures disappear from memory after a few seconds but can be reviewed through repeated exposure (?, ?, ?, ?). 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., ?, ?, ?, ?, ?). 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.

In contrast, forgetting/decay is critical to TW's 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 effects of 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 ?'s (?) 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 Figure 1), no forgetting at all (i.e., the input to each neurons was the cumulative sum of prior inputs; "Never" in Figure 1) and immediate forgetting (i.e., the activation decays immediately after presentation; "Immediate" in Figure 1). As shown in Figure 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 is it might, we believe that the merits of a psychological model should be evaluated by its empirical adequacy, and links between psychological parameters and neurobiological findings should be investigated empirically.

Regarding the importance for spreading activation for network performance, we agree, and, in their Section 2, EJ explained the role of spreading activation in detail.

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 derive closed-form expressions for the network dynamics (?, ?). Given that, in TW's model, interference and decay act on weights rather than activations, this problem might be somewhat reduced in their model, but it is still hard to evaluate the dynamic interplay of

first and higher order associations just based on the pattern fof weights.

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 formalism 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 the excitatory input (though the difference does not necessarily disappear; ?, ?). Given that weight changes are proportional to the product of the corresponding activations, connections between neurons with greater coactivation will still be strengthened to a greater extent. Again, we believe that targeted psychological experiments are necessary to gauge the empirical adequacy of activation-based vs. weight-based inhibition, but, 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.

Figure 2. Final weights after simulation of stream from ?'s (?). Mean (left) and maximal (right) weights for slow (top) and fast (bottom) forgetting rates.

TW also proposed some more specific criticisms of EJ's network. For example, TW argued that "it is not clear their [EJ's] model prevents excessive growth of connections" (p. ???). 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 ?'s (?) Experiment 2. As shown in Figure 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 we 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).

Related, TW questioned EJ's rationale for not varying their interference parameter (p. ???). 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 of external stimulation and noise, the activation change between times t and $t + 1$ is given by (exponential) decay (first term), spreading activation (second term) and inhibition (third term).

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

If the connectivity is relatively sparse (i.e., if there are many neurons), spreading activation will be limited, and only a relatively constant number of neurons will be active to deliver inhibitory input to all other neurons. In other words, the inhibitory input to each neuron will be relatively constant. Mathematically, the effect of constant inhibitory input is similar to that of decay, except that it is linear rather than exponential (though the specific functional form is likely more complex due to spreading activation). 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. However, TW's model confirms that EJ's results can be reproduced with different implementations.

In sum, both EJ and TW show that a memory-less correlational learning mechanism can account for Statistical Learning results, despite differences in implementation,

irrespective of whether decay and inhibition affect activations or weights. As a result, to the extent that Statistical Learning supports declarative memory formation for words, relevant evidence is still required.

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