Optimal storage capacity associative memories exhibit retrieval-induced forgetting

Andrew M. Saxe[1] & Kenneth A. Norman[2]

[1]Center for Brain Science, Harvard University [2]Department of Psychology, Princeton University

Overview

Retrieving a memory can, surprisingly, cause forgetting of related competitor memories, a phenomenon known as retrieval-induced forgetting. For example, after studying a list of category-exemplar pairs (``Fruit-Pear,'' ``Fruit-Apple'',..), partial practice of one target pair (``Fruit-Pe'') can cause forgetting of related competitor pairs (``Fruit-Apple''). A wealth of experiments have delimited four key features of this effect: *partial practice* yields retrieval-induced forgetting; *extra study* of the complete item (``Fruit-Pear'') yields no RIF despite equivalent target strengthening; *reversed practice* with incomplete category information (``F-Pear'') yields no RIF; and when present, the RIF effect can be elicited using *independent cues* (``Red-A'') rather than the specific cues used during learning (Norman et al., 2007; Anderson, 2003). These intricate findings pose a crucial challenge for theory: what sort of memory system might yield these effects, and why?

Here we develop a quantitative theory of retrieval-induced forgetting by deriving new exact solutions to the dynamics of learning for the generalized perceptron learning rule (GPLR) as it embeds memories in a binary recurrent neural network. These solutions yield closed-form expressions for the amount of RIF as a function of experimental parameters that agree with experiment. Hence the GPLR, which is known to attain optimal storage capacity in recurrent binary networks (Gardner, 1988), naturally exhibits retrieval-induced forgetting, suggesting that RIF is a hallmark of memory storage using a computationally optimal learning rule.

Retrieval-induced forgetting

Typical experiments consist of three phases.

Phase 1: Initial study

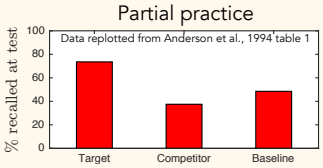
"Fruit-Pear"
"Fruit-Apple"
"Tool-Hammer"
...

Phase 2: Practice

Partial practice	*Reversed practice*	*Extra study*
"Fruit-Pe"	"Fr-Pear"	"Fruit-Pear"
RIF	No RIF	No RIF

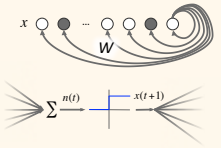
Phase 3: Test

Practiced cue	*Independent cue*
"Fruit-P"	"Red-A"
"Fruit-A"	
"Tool-H"	



Recurrent network model

- Memories stored in binary recurrent neural network
- All-to-all recurrent connections
- Patterns embedded as fixed points of network dynamics



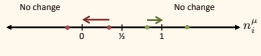
Generalized perceptron learning rule

- Introduced by Gardner, 1988

Net input to neuron i: $n_i^\mu = w_i x^\mu$

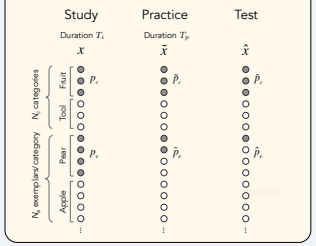
$$\Delta w_i = \nu \begin{cases} \mathrm{sgn}(x^\mu), & \text{if } x_i^\mu = 1 \text{ and } n_i^\mu < 1, \\ -\mathrm{sgn}(x^\mu), & \text{if } x_i^\mu = 0 \text{ and } n_i^\mu > 0, \\ 0 & \text{otherwise} \end{cases}$$

- Requires neurons to be correct by a nonzero margin (here ½)



- The GPLR is known to obtain optimal storage capacity of 2*N* (cf .14*N* for Hebbian learning) (Gardner, 1988).

Training paradigm and patterns



Exact solutions to the learning dynamics

We have found exact solutions for this setting as a function of $\nu, N_e, p_c, p_e, T_i, \tilde{p}_c, \tilde{p}_e, T_p, \hat{p}_c, \hat{p}_e$

c: target category unit
e: target exemplar unit
f: competitor exemplar unit

Study phase

$$n_c(t) = \min\left(1, \nu(N_e p_c + p_e) T_i\right)$$
$$n_e(t) = \min\left(1, \nu \frac{(N_e p_c + 1) p_e}{(N_e - 1) p_c + p_e} T_i\right)$$

Practice phase

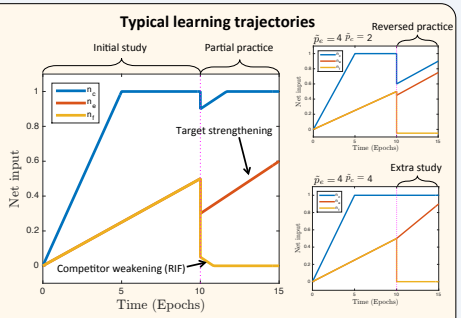
$$\tilde{n}_c(t) = a(T_i)(N_e \tilde{p}_c + \tilde{p}_e) + d_c(t)(\tilde{p}_c + \tilde{p}_e)$$
$$\tilde{n}_e(t) = b(T_i)\left[\frac{p_c \tilde{p}_e}{(N_e - 1)p_c + p_e}\right] + d_e(t)(\tilde{p}_c + \tilde{p}_e)$$
$$\tilde{n}_f(t) = b(T_i)\left[\frac{\tilde{p}_c p_e - p_c \tilde{p}_e}{(N_e - 1)p_c + p_e}\right] + d_f(t)(\tilde{p}_c + \tilde{p}_e)$$

Test phase

$$\Delta\hat{n}_c = d_c(T_i + T_p)\left(\min(\tilde{p}_c, \hat{p}_c) + \min(\tilde{p}_e, \hat{p}_e)\right)$$
$$\Delta\hat{n}_e = d_e(T_i + T_p)\left(\min(\tilde{p}_c, \hat{p}_c) + \min(\tilde{p}_e, \hat{p}_e)\right)$$
$$\text{RIF: } \Delta\hat{n}_f = d_f(T_i + T_p)\min(\tilde{p}_c, \hat{p}_c)$$

Functions $a(t), b(t)$, and $d_{c,e,f}(t)$ omitted due to space

Typical learning trajectories

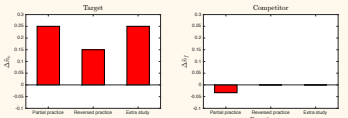


Parameters: $N_e = p_e = p_c = \tilde{p}_c = \hat{p}_c = 4, \tilde{p}_e = 2, \hat{p}_e = 1, T_i = 10, T_p = 5, \nu = .01$

Effect of practice type

Key results

- Consistent with experiment, partial practice yields RIF
- Reversed/extra study yield no RIF, despite substantial target strengthening

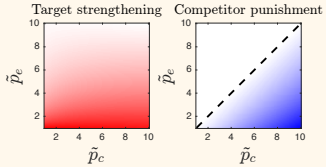


$N_e = p_e = p_c = \hat{p}_c = 4, \hat{p}_e = 1, T_i = 10, T_p = 5.$

Asymptotic RIF

- If training is allowed to proceed to convergence $(T_i, T_p \to \infty)$, then expressions simplify:

$$\Delta\hat{n}_e = \left(\frac{1}{\tilde{p}_e + p_e} - \frac{p_e \tilde{p}_e + p_e \hat{p}_e + (N_e - 1)p_e \hat{p}_e}{(N_e p_e + p_e)(\tilde{p}_c + \tilde{p}_e)p_e}\right)\left(\min(\tilde{p}_c, \hat{p}_c) + \min(\tilde{p}_e, \hat{p}_e)\right)$$
$$\Delta\hat{n}_f = \min\left(0, -\frac{\tilde{p}_c p_e - p_c \tilde{p}_e}{(N_e p_e + p_e)(\tilde{p}_c + \tilde{p}_e)p_e}\right)\min(\tilde{p}_c, \hat{p}_c)$$

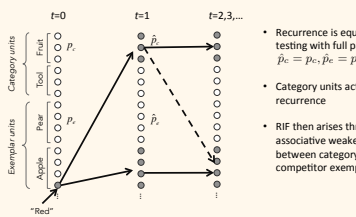


- No RIF is observed when less category information than exemplar information is presented

Other parameters are $N_e = p_e = p_c = \hat{p}_c = 10, \hat{p}_e = 1.$

RIF with independent cues

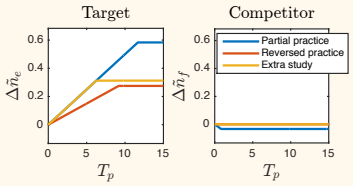
- A key finding of experimental literature is that RIF can be observed even with cues not used during training or practice (E.g., study/practice "Fruit-Apple", test "Red-A")
- Recursive dynamics enable independent cue results



- Recurrence is equivalent to testing with full pattern, $\tilde{p}_c = p_c, \tilde{p}_e = p_e$
- Category units activate due to recurrence
- RIF then arises through associative weakening between category and competitor exemplar units
- Model prediction: Independent-cue RIF is equivalent to practiced-cue RIF

RIF with varying amounts of practice

- Target strengthening persists while RIF rapidly plateaus with practice



Conclusions

- Theory points to a computational rationale for RIF: phenomena relating to RIF are natural consequences of memory storage using a computationally optimal learning rule
- Makes quantitative, testable predictions for the exact degree of RIF as a function of experimental parameters
- First analytical model to capture the basic phenomenology of RIF
- Links neural plasticity to high level psychological phenomenon, showing how a network of neurons with local learning could combine to yield the behavioral patterns of RIF
- By virtue of its neural formulation, the model may address more recent neural data (Poppenk & Norman, 2014; Wimber et al., 2015)
- Solution methods employed may be generalizable to other emerging RIF phenomena such as reverse RIF, integration, and differentiation (Hulbert & Norman, 2014).

Support: A.M.S. is supported by a Swartz Postdoctoral Fellowship.