Optimal storage capacity associative memories exhibit retrieval-induced forgetting

Andrew M. Saxe\textsuperscript{1} & Kenneth A. Norman\textsuperscript{2}

\textsuperscript{1}Center for Brain Science, Harvard University \textsuperscript{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"), a person is likely to forget the completion pairs ("Pear-Fruit," "Apple-Fruit"). A wealth of experiments have highlighted four key factors of this effect: partial pairings yield retrieval-induced forgetting; omit one of the elements in any of those pairs ("Pear-Fruit") yields no RIF despite the presence of a full completion pair; introduction of a new memory ("F. Pear") yields no RIF, and when present, the RIF effect can be abolished using independent cues. The specific cue used during learning (Norman \textit{et al.}, 1997; Anderson, 2003). These indicate feedbacks on a neural 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 perception-learning rule (GPLR) as it manifests memory in a highly recurrent neural network. These solutions yield closed-form expressions for the amount of RIF as a function of experimental parameters, and show that the GPLR is a hallmark of memory systems using a computationally optimal learning rule.

### Recurrent network model

#### Generalized perception learning rule

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

\[
\Delta x_i(t) = \left( - \alpha x_i(t) + \sum_{j} w_{ij} x_j(t) + \sum_{k} p_k(t) \right) \delta(t)
\]

where $\alpha$ is the learning rate, $x_i(t)$ is the state of unit $i$, $w_{ij}$ is the weight between units $i$ and $j$, and $p_k(t)$ are the inputs to the network.

- Dependences on multiple experimental parameters:
  - Interconnection strengths
  - Input strengths
  - Learning rate

### Exact solutions to the learning dynamics

We have found exact solutions for this setting as a function of $\alpha$, $\beta$, $\gamma$, $\delta$, $\epsilon$, $\zeta$, $\eta$, $\theta$, and $\phi$. The GPLR is known to obtain optimal storage capacity of 2N (C. G. K. Fisher learning (Gardner, 1988)).

### Effect of practice type

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

### 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 parameters:
- First analytical model to capture the basic phenomenology of RIF
- Links neural plasticity to high-level psychological phenomena, showing how a network of neurons with local learning could combine to yield the behavioral patterns of RIF
- By virtue of its novel formulation, the model may address more recent neural data (Popper & Norman, 2010; Winter et al., 2010)
- Solutions methods employed may be generalizable to other emerging RIF phenomena such as reverse RIF, inhibition, and differentiation (Insdorf & Norman, 2014).

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