

Balancing Long-Term Reinforcement and Short-Term Inhibition

Christian Lebiere (cl@cmu.edu)

Carnegie Mellon University

Bradley Best (bjbest@adcogsys.com)

Adaptive Cognitive Systems

Outline

- Rational analysis of memory
- Computational implementation
- Long-term impact of short-term effects
- Revisiting the environment
- Combining reinforcement and inhibition
- Internal dynamics mirror environment

Rational Analysis of Memory

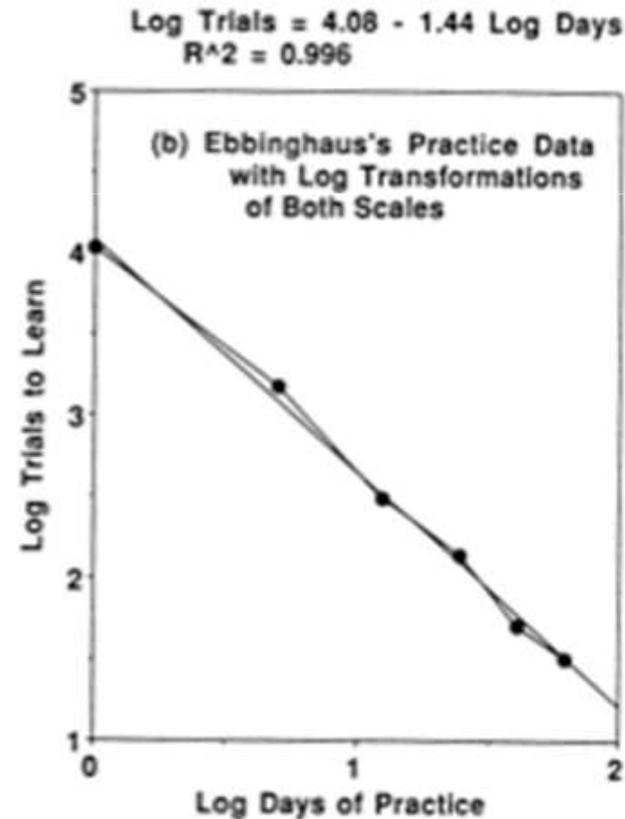
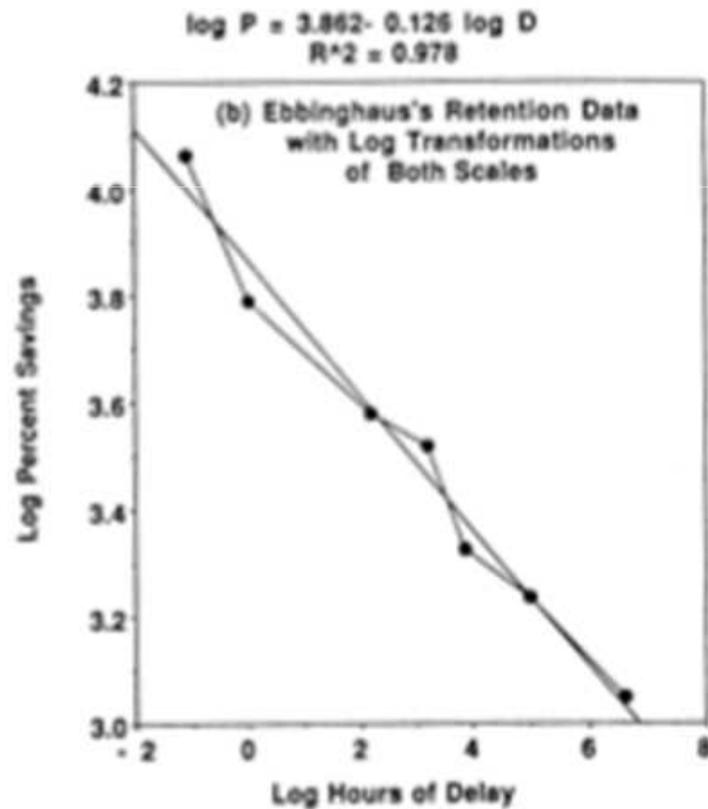
- Human memory has adapted through evolution to the structure of its environment (Anderson & Schooler, 1991)
 - Frequency effects
 - Recency effects
 - Spacing effects
- Even apparent failures serve a functional role
 - Forgetting as a way of managing scale of memory demands
- Given resource constraints on long-term memory, optimal behavior is making most available the memories that are most likely to be needed

Power Law of Memory

Retention and practice functions for range of measures

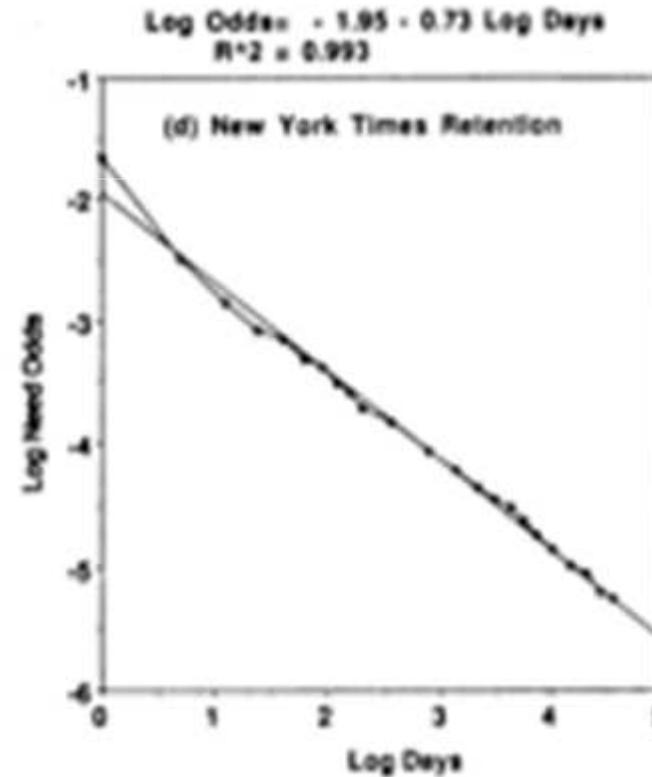
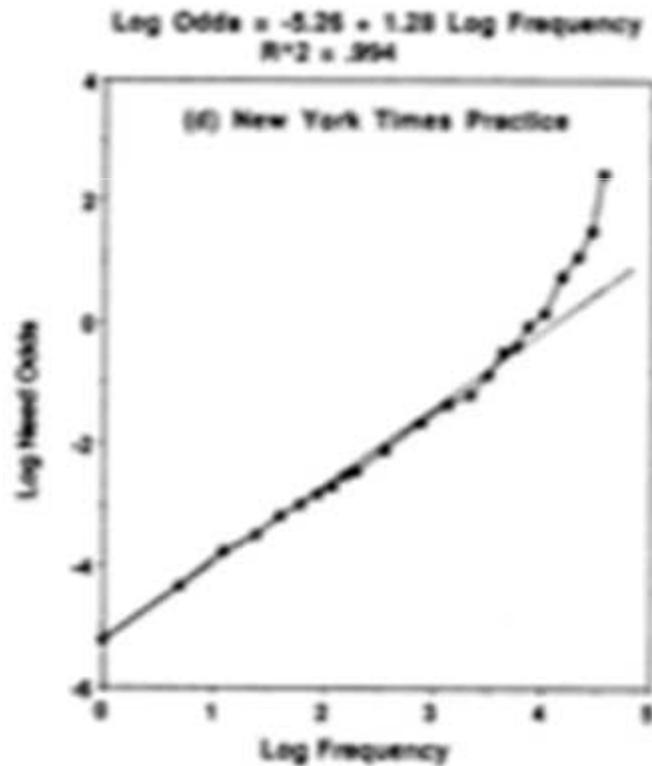
$$P = AT^{-b}$$

$$\log P = \log A - b \log T$$



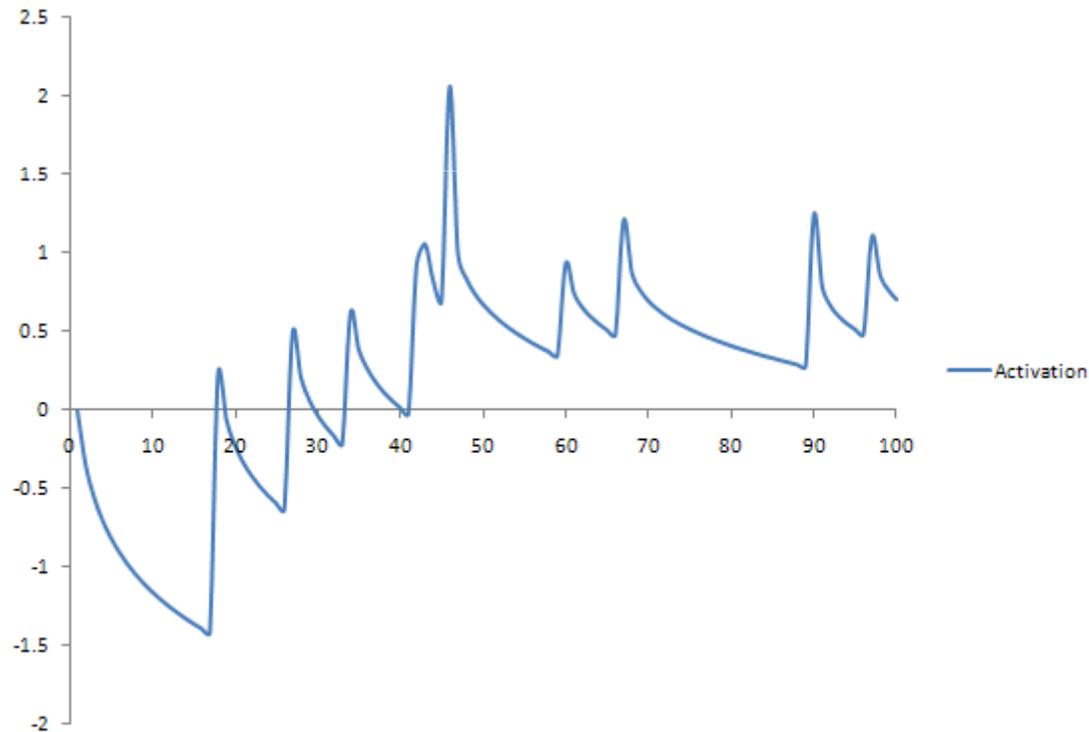
Power Law of the Environment

Need Odds follow power law in human environments:
NYT headlines, CHILDES speech db, email addresses



Computational Implementation

- Delay and Practice are roughly additive effects
- Activation as log odds modulates recall and latency



$$\frac{Prob}{1-Prob} = Odds = a \sum_1^n t_k^{-d}$$

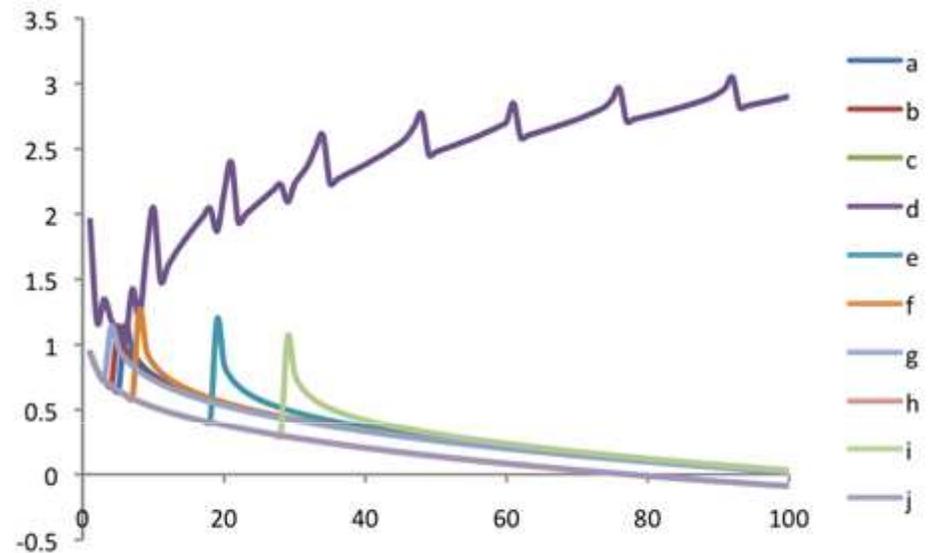
$$A_i = B + \ln \left(\sum_1^n t_k^{-d} \right)$$

In The Short Run

- “A power function implies that the performance measure will go to infinity as time goes to zero.” p.398
- “Power functions for forgetting tend to be obtained when we use measures that do not have upper bounds or do not approach their upper bounds.” p.398
- “ Power functions seem to describe memory performance from a few seconds to years.” p.398
- What are the long-term implications if performance measures grow arbitrarily large under a few seconds?

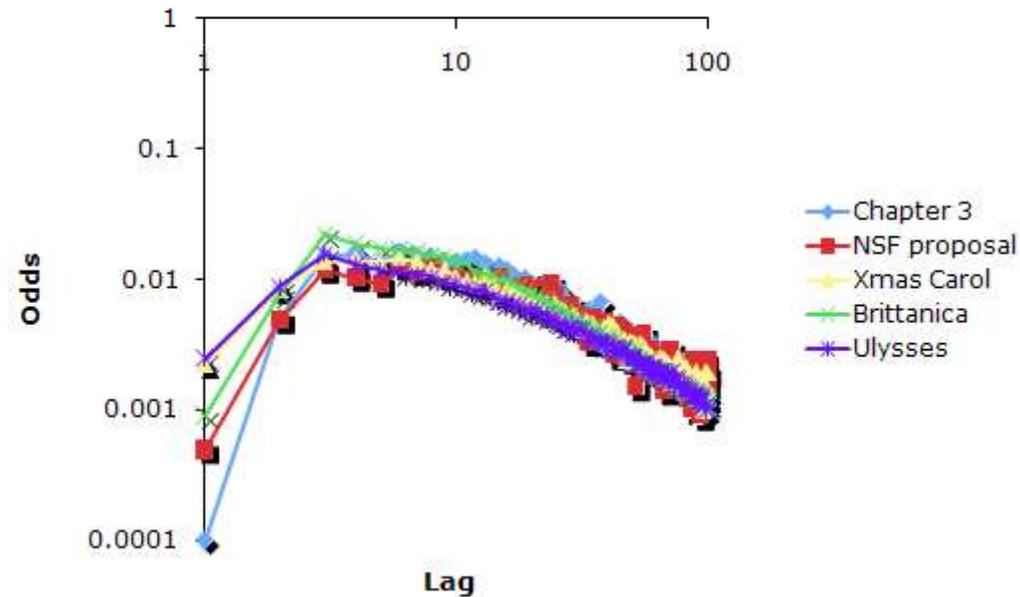
Winner-Take-All Dynamics

- Reinforcement leads to feedback loop
- Strong conditions on retrieval can help...
- But they detract from the use of activation
- Need partial matching and under-constrained retrievals
- Can be controlled by modeler with tags or finsts
- Metacognitive knowledge hard to specify at high-level
- Need robust, general models of open-ended behavior



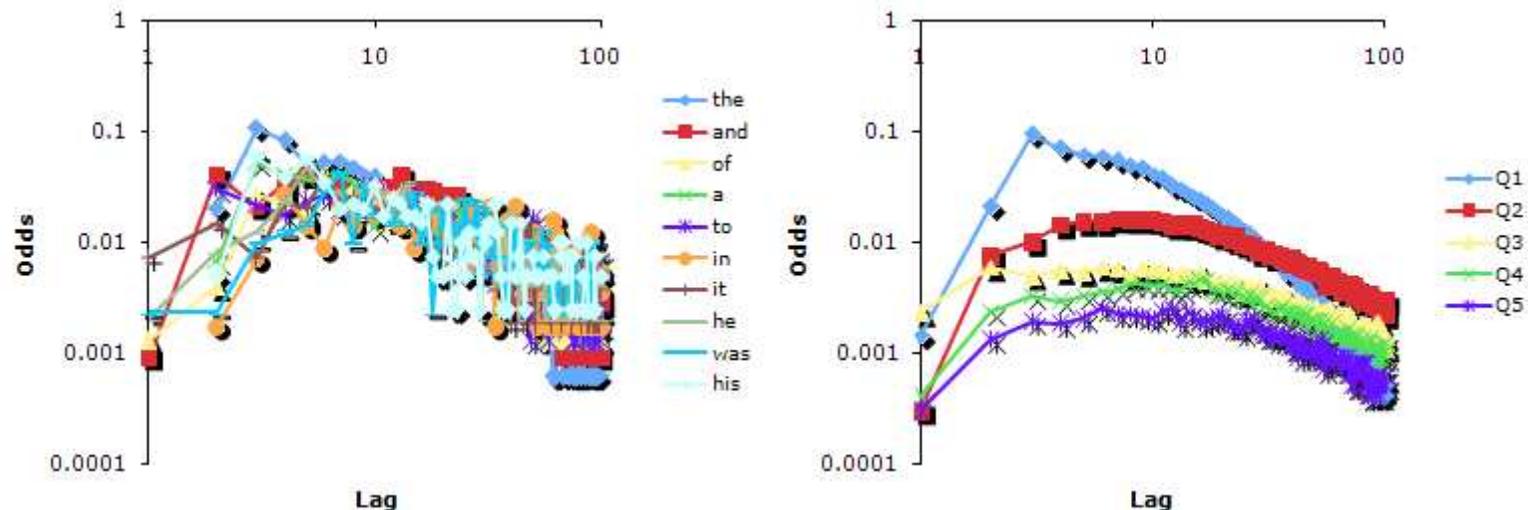
Back to the Environment

- Language as preeminent sequential structured envt
- Short-term depression in need odds at small lags
- Consistent across very different text corpora



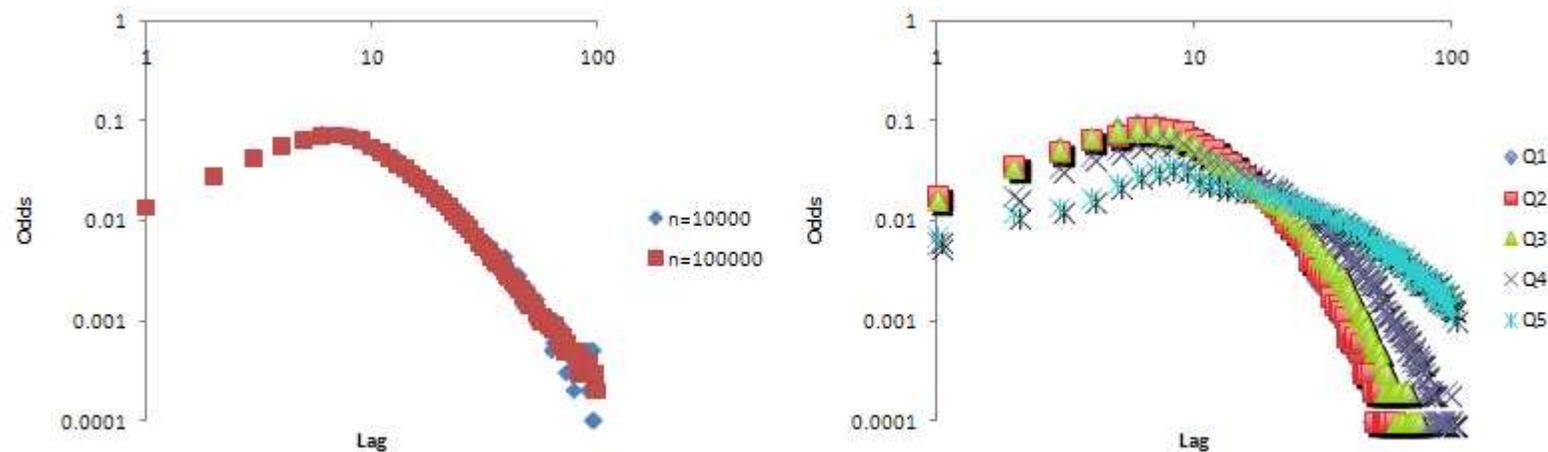
Words and Groups

- The pattern holds for highest frequency words
- Additive effect is observed at all levels of frequency
- Peak in odds boost consistent \sim lag 5-10 for all groups



Access to Arithmetic Facts

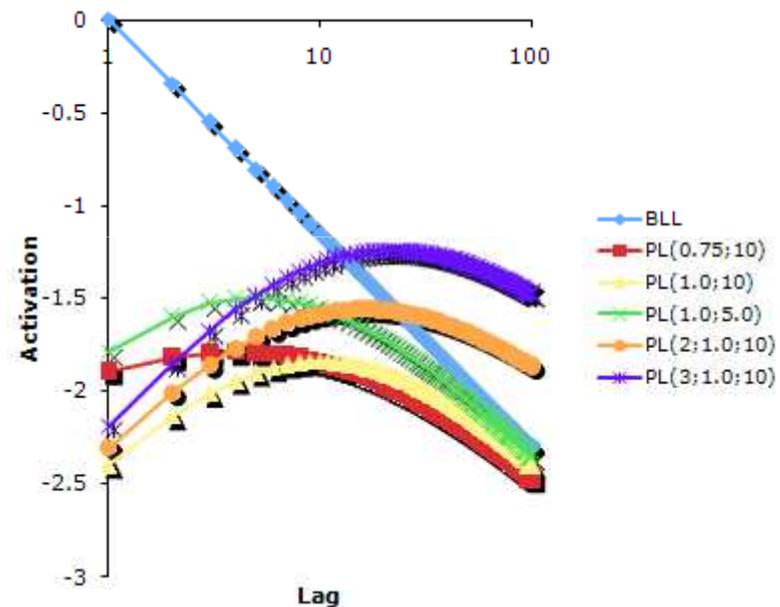
- As in original analysis, effect must hold across domains
- Arithmetic domain generated from validated model
- Pattern holds including peak and size of inhibition
- Apply to other environments, e.g. web, physical spaces



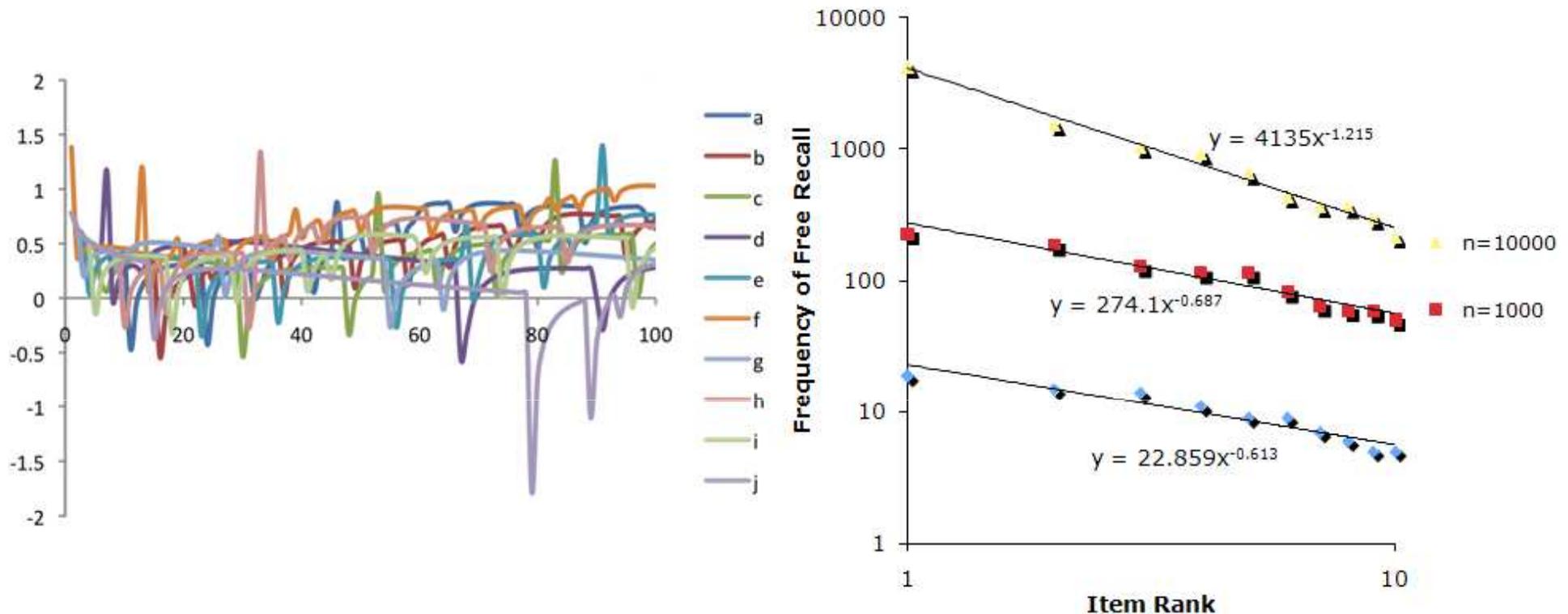
Activation Inhibition

$$B_i = \log \sum_{j=1}^n t_j^{-d} - \log \left(1 + \left(\frac{t_n}{t_s} \right)^{-d_s} \right)$$

- Parameters:
 - Inhibition scale t_s – controls period to peak reinforcement
 - Inhibition decay t_d – specifies magnitude of inhibition effect
- Additive term integrates with other terms of activation equation: noise, spreading association, partial matching



Emergent Robustness



- Soft inhibition differs from pathological behavior of the default version and from the hard and fixed round-robin of the first version
- Running the retrieval mechanism *unsupervised* leads to the gradual emergence of an internal power law distribution *internally*

Discussion

- Biological implementation of short-term inhibition
 - Short-term depotentiation?
- Architectural implications of short-term inhibition
 - Working memory and refraction
- Contribution of other cognitive factors in task
 - Grammatical rules, base-10 systems
 - Could have evolved in response to cognitive limitations
- Integrate combination of environmental, neural and behavioral constraints in cognitive architectures