Continual learning, replay and consolidation in a forgetful memory network model

Nimrod Shaham[1], Jay Chandra[1], Gabriel Kreiman[1,2] and Haim Sompolinsky[1,3]

[1]Harvard University Center for Brain Science, [2]Harvard medical school, [3]Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem

Introduction

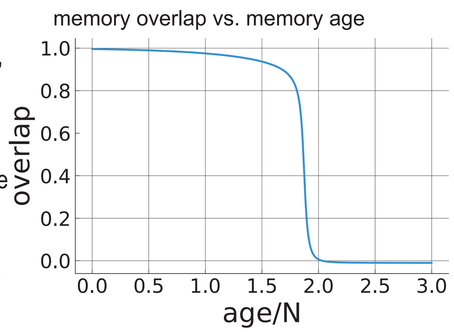
- The brain continually stores memories, some of them stay for a lifetime, some forgotten.
- Empirical memory studies show consolidation effects, presumably through memory replay, and slow decaying forgetting curves.
- Palimpsestic models incorporating weight decay allow for lifelong learning, but predict sharp and unrealistic forgetting curves.
- We introduce stochastic nonlinear replay model, which provides a framework for studying forgetting and replay effects on lifelong learning.
- The model leads to replay-dependent memory consolidation, with long memory lifetimes and slowly decaying forgetting curves.

Sparse palimpsest model

We first study a sparse extension of the palimpsest model of Nadal et al., (1986), Mezard et al., (1986): Attractor network of size N, and a large number of sparse binary memories (sparseness f) a la Tsodyks and Feigelman (1987) gives rise to exponentially decaying memory efficacies with age n.

$$J_{ij} = \frac{1}{Nf(1-f)} \sum_n A(n)(\xi_i^n - f)(\xi_j^n - f)$$

$$A(n) = e^{-n/\tau}$$



The activity overlap with a memory state decays with age deterministically. Retrievability (large enough overlap) sharply falls to zero at a critical age.

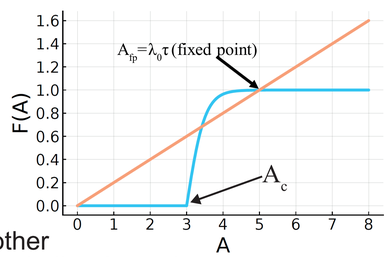
Stochastic nonlinear replay model

- Memories are replayed at random times and undergo consolidation in the form of Hebbian plasticity after each replay.
- The rate of replay for a given memory is proportional to its basin of attraction, which is a function of its current efficacy. Importantly - a memory is replayed only if it is currently an attractor of the network's dynamics.

$$J_{ij}(t) \sim \sum_n A_n(t)(\xi_i^n - f)(\xi_j^n - f)$$

$$\frac{dA_n}{dt} = -\frac{1}{2N\tau} A_n + \eta_n(t), \quad \eta_n(t) = \sum_{\{t_n\}} \delta(t - t_n)$$

$$\langle \eta_n(t) \rangle \equiv \lambda(A(t)) \equiv \frac{\lambda_0}{2N} F(A/\Delta) \qquad \Delta^2 = f \cdot \frac{1}{N} \sum_n A_n^2$$

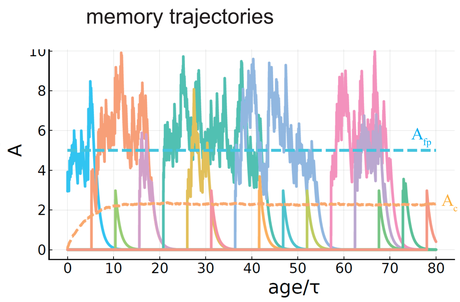


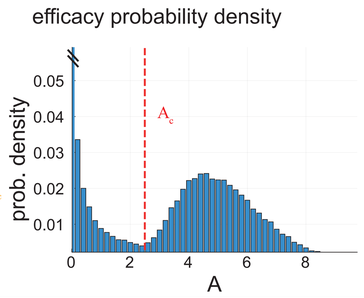
Δ- the std of local fields in a memory state, due to the interference with all other memories. Determined self consistently in the dynamics.

A_c - the critical efficacy for an attractor, prop. to Δ. $A(0)$ - initial memory efficacy, assumed to be >1, and might be larger or smaller than A_{fp}. Mean field efficacy dynamics: three fixed points. Stable ones: $A=0$, $A=\lambda_0\tau=A_{fp}$.

Dynamics of memory efficacies

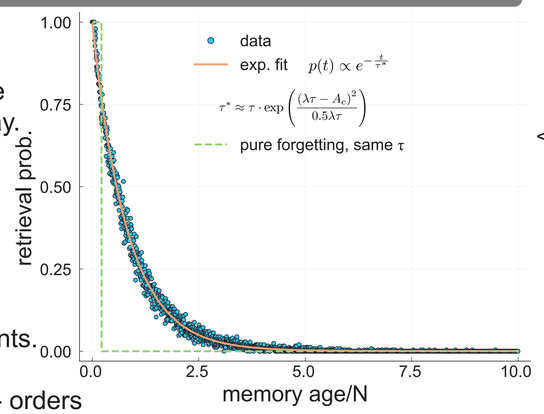
- At first memories tend to go towards A_{fp}, and efficacies grow.
- Next, due to the stochastic nature of replay, memories might fall toward zero. The decay time is very long compared to τ.





Forgetting curve

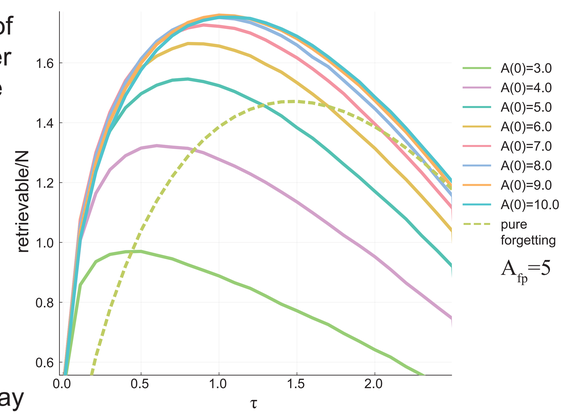
- Retrieval decays gracefully with age, in contrast to the critical age in the model of pure decay.
- Memories much older than the critical age of pure decay have non-zero retrieval probability.
- The decay is exponential - consistent with behavioral experiments.
- The slow decay characteristic time scale - orders of magnitude larger than the forgetting time scale - results from a long mean first passage time of the efficacies from the fixed point to the critical efficacy (see inset equation).



Capacity

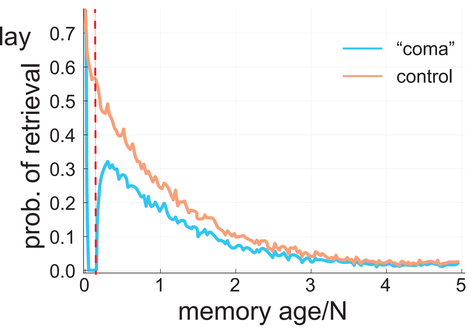
Capacity: total number of retrievable memories per neuron. Depends on the decay rate, initial efficacies and averaged replay rate.

- For fast decay rate, replay has significant advantage over pure forgetting.
- For slow decay rate, the pure decay and replay models converge - for slow enough decay rate, the noise (and thus A_c) crosses $A(0)$, and then there are no replays, as in pure decay.



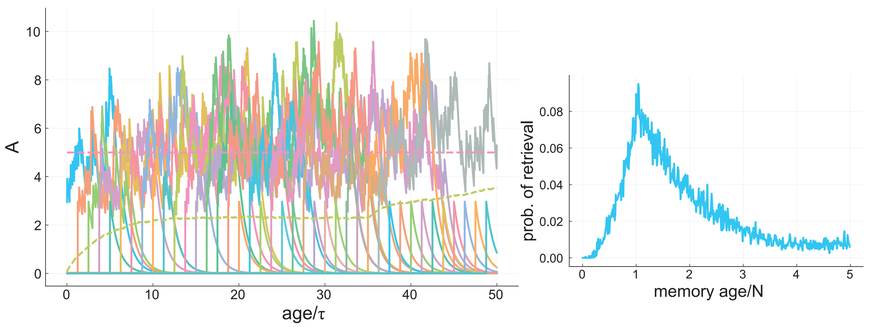
Memory pathologies

"Coma": transient pause in replay and entry of new memories. After recovery, the memories learnt just before the coma onset are impaired due to insufficient consolidation. This gives rise to a non-monotonic forgetting curve.



Network damage, lesions: Random synaptic/ cell death.

- Abrupt and transient event ("TIA"): effect similar to "coma".
- "Dementia": gradual damage, linear increase in fraction of dead synapses. The ability to consolidate new memories is damaged, while some of the old consolidated memories are spared.



Conclusions

- Pure decay palimpsest models exhibit catastrophic forgetting as a function of memory ages.
- We introduce a consolidation model based on stochastic nonlinear replay where replay rates of individual memories depend on the size of their basins of attraction.
- Our model generates graceful forgetting, with forgetting curves shape consistent with experiments, and with long time constants.
- Initial memory efficacies play a major role, while the pure forgetting model is indifferent to them.
- Model capacity is similar or larger than that of pure decay models.
- Model forgetting curves exhibit interesting non-monotonic memory deficits under scenarios mimicking clinical pathologies of human memory.