Funded by the U.S. National Science Foundation and the Simons Foundation

Finding robust memories through representational drift

Maanasa Natrajan, James E. Fitzgerald

Memories are believed to be stored in patterns of synaptic connections, physical links that allow neurons to coordinate their activity. Rather than acting alone, neurons form coordinated groups called neural ensembles, which encode and retrieve complex information. Learning occurs when the connection strengths between neurons (synaptic weights) change, but modifying these connections can interfere with previously stored memories. As a result, the brain faces a fundamental tradeoff between plasticity, the ability to learn new information, and stability, the preservation of existing memories.

Surprisingly, the neural ensembles underlying memory continue to change even in the absence of apparent learning or forgetting, a phenomenom known as representational drift. Theoretical studies suggest that multiple neural representations can encode the same memory. However, it remains unclear whether the representations explored through drift offer unique advantages over those reached through initial learning.

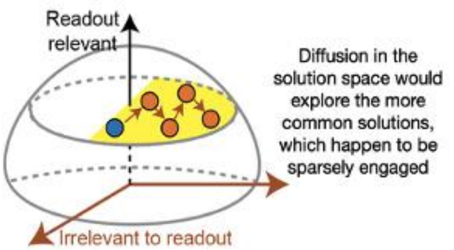

Figure 1: We mathematically model representational drift as bounded diffusion (contained wandering) within the solution space of synaptic weights. The space is defined as the manifold of all synaptic weight matrices that exactly satisfy a specific set of input-output mappings in a neural network model.

To investigate this, we characterized the non-linear “soulution space” of a memory, the set of all synaptic configurations that produce the same input-output behavior in a neural ensemble. By modeling representational drift as a gradual diffusion within this space, we were able to distinguish drift from learning and forgetting (Figure 1). We found that representational drift uncovers noise-robust representations that reliably preserve information despite background variability. These representations are sparsely engaged, with many neurons either inactive or saturated, a disengaged state that confers robustness against weight perturbations from noise or continual learning (Figure 2). These robust solutions are abundant within the solution space and are entropically favored by drift, but their lack of gradients makes them difficult to learn and non-conducive to further learning.

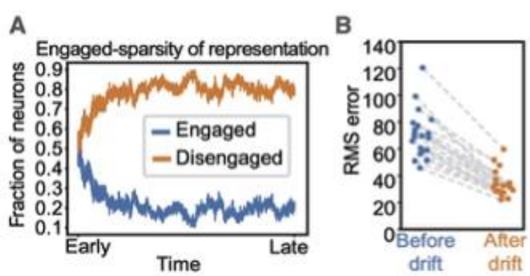

Figure 2: Representational drift leads to sparsely engaged representations that are robust to weight perturbations. (A) This panel shows the fraction of engaged and disengaged (inactive or saturated) neurons over time. Drift triggers an initial drop in the fraction of engaged neurons followed by stable maintenance. (B) Root-mean squared (RMS) error due to weight perturbations before drift (blue) and after drift (orange).

Importantly, no single representation is both maximally robust and optimally suited for learning. To resolve this tradeoff, we introduce an allocation procedure that allows the system to dynamically shift from robust, stability-preserving states to more flexible, learning-conducive ones. Together, representational drift and allocation enable the brain to balance stability and plasticity over time, providing a new framework for understanding how memories remain reliable while synapses and neural ensembles continue to change.

National Institute for Theory and Mathematics in Biology

Funded by the U.S. National Science Foundation and the Simons Foundation

Reference: https://doi.org/10.1073/pnas.2500077122. This research was supported in part by grants from the NSF (DMS-2235451) and Simons Foundation (MP-TMPS-00005320).