National Institute for Theory and Mathematics in Biology
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 synapses and retrieved through the reactivation of neural ensembles. Learning alters synaptic weights, which can interfere with previously stored memories that share the same synapses, creating a tradeoff between plasticity and stability. Interestingly, the neural ensembles underlying memory exhibit significant dynamics, even in stable environments and without apparent learning or forgetting - a phenomenon known as representational drift. Theoretical studies have suggested that multiple neural representations can correspond to a memory, with post-learning exploration of these representation solutions driving drift. However, it remains unclear whether representations explored through drift differ from those learned or offer unique advantages.

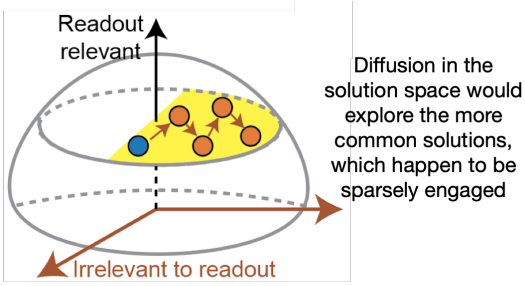

Fig. 1. We mathematically model representational drift as bounded diffusion in the solution space of synaptic weights. This solution space is defined to be the set of synaptic weight matrices that exactly memorize a set of specified input-output pairs in a neural network model.

Here we show that representational drift uncovers noise-robust representations that are otherwise difficult to learn. We first define the non-linear solution space manifold of synaptic weights for a fixed input-output mapping, which allows us to disentangle drift from learning and forgetting and simulate representational drift as diffusion within this manifold (Figure 1). Solutions explored by drift are sparsely engaged, i.e. have many inactive and saturated neurons (collectively termed disengaged), making them robust to weight perturbations due to noise or continual learning (Figure 2). Such solutions are prevalent and entropically favored by drift, but their lack of gradients makes them difficult to learn and non-conducive to further learning. To overcome this, we introduce an allocation procedure that selectively shifts representations for new information into a learning-conducive regime. By combining allocation with drift, we resolve the tradeoff between learnability and robustness.

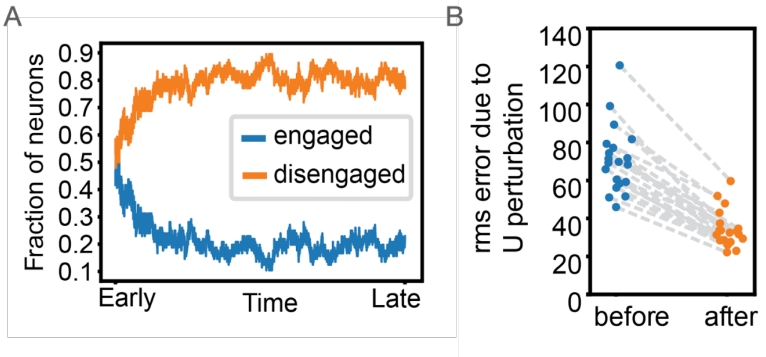

Fig. 2. Representational drift leads to sparsely engaged representations that are robust to weight perturbations. (A) Fraction of engaged and disengaged (inactive and saturated) neurons over time during drift showing an initial drop in the fraction of engaged neurons followed by stable maintenance. (B) Root-measure squared error due to weight perturbations before drift (blue) and after drift (orange)

Reference: *bioRxiv* doi/10.1101/2024.12.19.629245. This research was supported in part by grants from the NSF (DMS-2235451) and Simons Foundation (MP-TMPS-00005320).