Coordinated drift of receptive fields during noisy representation learning

Shanshan Qin[1,2], Shiva Farashahi[3,†], David Lipshutz[3,†], Anirvan M. Sengupta[3,4], Dmitri B. Chklovskii[3,5], and Cengiz Pehlevan [*1,2]

[1] *John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, U.S.A*
[2] *Center for Brain Science, Harvard University,Cambridge, MA 02138, U.S.A*
[3] *Center for Computational Neuroscience, Flatiron Institute, New York, NY 10010, U.S.A*
[4] *Department of Physics and Astronomy, Rutgers University, New Brunswick, NJ 08901, U.S.A*
[5] *NYU Langone Medical Center, New York, NY 10016, U.S.A*

Abstract

Long-term memories and learned behavior are conventionally associated with stable neuronal representations. However, recent experiments showed that neural population codes in many brain areas continuously change even when animals have fully learned and stably perform their tasks. This representational "drift" naturally leads to questions about its causes, dynamics, and functions. Here, we explore the hypothesis that neural representations optimize a representational objective with a degenerate solution space, and noisy synaptic updates drive the network to explore this (near-)optimal space causing representational drift. We illustrate this idea in simple, biologically plausible Hebbian/anti-Hebbian network models of representation learning, which optimize similarity matching objectives, and, when neural outputs are constrained to be nonnegative, learn localized receptive fields (RFs) that tile the stimulus manifold. We find that the drifting RFs of individual neurons can be characterized by a coordinated random walk, with the effective diffusion constants depending on various parameters such as learning rate, noise amplitude, and input statistics. Despite such drift, the representational similarity of population codes is stable over time. Our model recapitulates recent experimental observations in hippocampus and posterior parietal cortex, and makes testable predictions that can be probed in future experiments.

[*]Corresponding author. E-mail: cpehlevan@seas.harvard.edu
[†]These authors contributed equally to this work

Introduction

Memories and learned behavior can be stable for a long time. We can recall vividly the memory of events that happened years ago. Motor skills, such as riding a bike, once learned, can last life-long even without further practice. A natural question is then whether stable task performance and memories are related to stable neuronal representations.

Recent technical advances in electrophysiology and optical imaging enabled researchers to address this question by studying the long-term dynamics of neural population activity in awake behaving animals (Katlowitz, Picardo, and Long 2018; Li et al. 2017; Luo et al. 2020; Schoonover et al. 2021; Ulivi et al. 2019; Y. Ziv et al. 2013). A number of these experiments have shown that neuronal activities in cortical areas that are essential for specific tasks undergo continuous reorganization even after the animals have fully learned their tasks, a phenomenon termed "representational drift" (Mau, Hasselmo, and Cai 2020; Rule, O'Leary, and Harvey 2019). For instance, in sensorimotor tasks, neuronal representations in the posterior parietal cortex (PPC) of mice change across days while the performance of animals remain stable and high (Driscoll et al. 2017). Place fields of individual place cells in CA1 region of hippocampus drift over days and weeks even when the animals remain in the same familiar environment (Gonzalez et al. 2019; Lee et al. 2020; Y. Ziv et al. 2013). Individual neurons in the primary motor cortex and supplementary motor cortex show unstable tuning while animals perform highly stereotyped motor tasks (Rokni et al. 2007) (but see (Chestek et al. 2007; Gallego et al. 2020)). Representational drift has been observed even in primary sensory cortices, such as mouse visual cortex (Deitch, Rubin, and Y. Ziv 2021; Marks and Goard 2021) and piriform cortex (Schoonover et al. 2021). The ubiquity of representational drift raises several important questions: What is the underlying mechanism of such drift? How can neural circuits generate stable coding in the presence of continuous drift? What are the dynamics of representational drift?

To address these issues, we consider a setting where a neural population learns to represent stimuli in a way that optimizes a representational objective. Such a normative account of sensory representations is common in neuroscience (Atick and Redlich 1992; Attneave 1954; Barlow 1961; Chalk, Marre, and Tkacik 2018; Hateren 1992; Olshausen and Field 1997; Pehlevan, Hu, and Chklovskii 2015; Rao and Ballard 1999; Srinivasan, Laughlin, and Dubs 1982). Furthermore, the representational objective we consider has many solutions, consistent with the notion that there are many optimal neural representations of input stimuli. Based on the observations that synapses in the cortex are highly dynamic (Attardo, Fitzgerald, and Schnitzer 2015; Hazan and N. E. Ziv 2020; Rumpel and Triesch 2016), we hypothesize that noisy synaptic updates during learning will drive the network to explore the synaptic weight space that corresponds to (near-)optimal neural representations. In other words, the neural

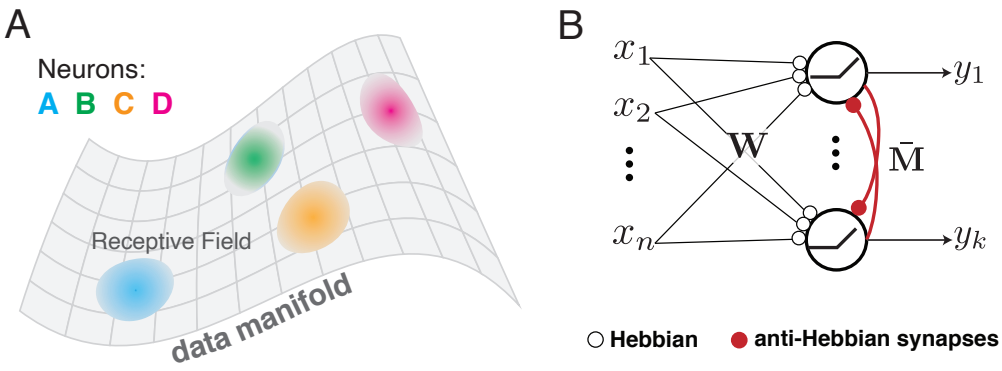

Figure 1: (A) Illustration of localized receptive fields that tile the data manifold. (B) Hebbian/anti-Hebbian network with nonnegative neural activity can learn localized receptive fields.

representation will drift within the space of optimal representations.

To test this hypothesis, we focus on a well-studied biologically plausible network for representation learning: the Hebbian/anti-Hebbian network (Földiak 1990; Pehlevan and Chklovskii 2019) (Fig. 1B). These networks optimize similarity matching objectives which exhibit a degeneracy of optimal representational solutions (Pehlevan, Sengupta, and Chklovskii 2018), all of which share the same representational similarity matrix (Kriegeskorte, Mur, and Bandettini 2008). Hebbian/anti-Hebbian networks have also been shown to learn localized RFs that tile the input data manifold (Sengupta et al. 2018), hence they can be used as simplified models for brain areas where neurons have localized receptive fields (RFs), such as hippocampal place cells and neurons in PPC (Driscoll et al. 2017; Gonzalez et al. 2019; Y. Ziv et al. 2013). In these systems, population of neurons with different localized RFs generally tile the parameter space they encode (Fig. 1A). The simplicity and mathematical-tractability make Hebbian/anti-Hebbian networks an excellent starting point to elucidate the mechanism and properties of representational drift due to noise in synaptic updates.

By numerical and analytical methods, we find that while the RFs of individual neurons change significantly over time, the representational similarity of population codes is stable.We show that the drift dynamics of individual RFs can be largely captured by a random walk on the data manifold, with the effective diffusion constant depending on noise amplitude, learning rate and other model parameters such as the number of output neurons. At the population level, the drifting RFs are coordinated in a way that preserves representational similarity. Our model accounts for many of the recent experimental observations in the hippocampus and PPC, and makes testable predictions. Overall, our results show how optimal representation learning and noise can lead to representational drift while maintaining representational similarity.

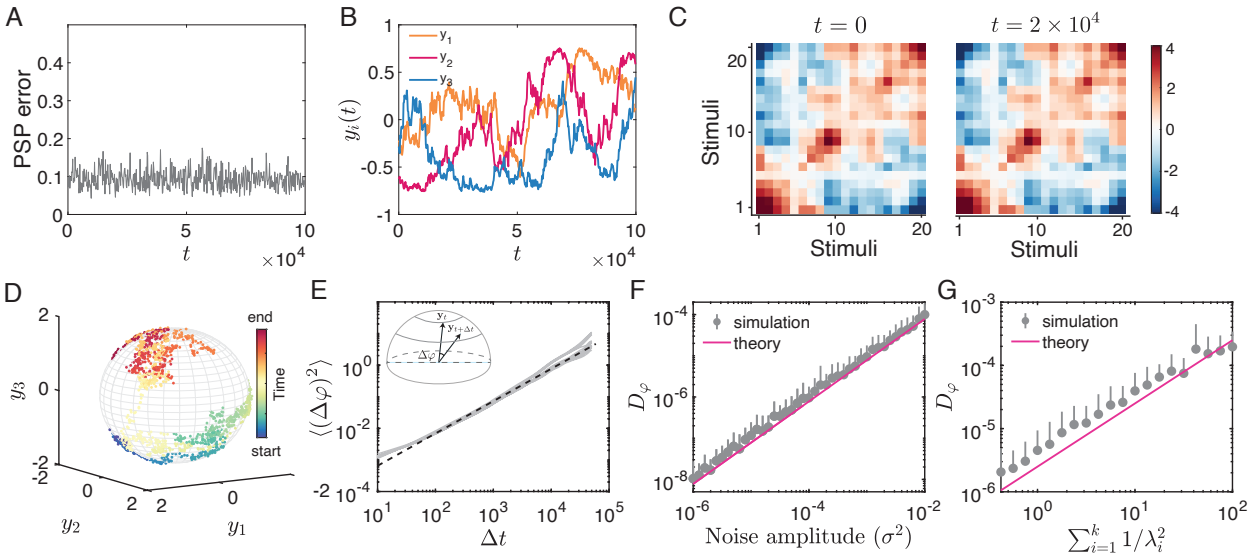

Figure 2: Drift dynamics in principal subspace projection (PSP) task. In the simulation, each input $\mathbf{x}_t \in \mathbf{R}^{10}$ is drawn independently from a joint Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{C})$. The first three eigenvalues of the covariance matrix $\mathbf{C}$ are: 4.5, 3.5, 1 and the rest are 0.1. A Hebbian/anti-Hebbian network learns to project this input to $k = 3$ dimensions. (A) PSP error remains stable after the task has been learned even with noisy synapse update. (B) The learned representation of an example input continuously changes due to noisy updates. Shown are the 3 component of $\mathbf{y}(t)$. (C) Pairwise similarity between learned representations are stable over time, as shown by the almost identical similarity matrices at $t = 0$ (left) and $t = 2 \times 10^4$ (right). (D) Drifting representation as a random walk on a "sphere", showing the representation of a single sample $\mathbf{y}_t$ over time . Color codes for different time steps. (E) Estimating rotational diffusion constant D_φ from mean squared angular displacement (MSAD). Grey lines are MSAD estimated based on individual representation trajectory $\mathbf{y}(t)$. The dashed line is a linear fit between $\langle(\Delta\varphi)^2\rangle \equiv \langle(\varphi(t+\Delta t) - \varphi(t))^2\rangle$ and Δt to estimate the rotational diffusion constant. Inset: illustration of $\Delta\varphi$. (F) Relationship between D_φ and noise amplitude σ^2. Symbol with error bars are numerical simulation, and the solid line is the theory Eq.5. (G) Dependence of D_φ on the eigenspectrum $\{\lambda_i\}$ of the input covariance matrix. Error bars represent standard deviation, only one side is shown to reduce cluttering. In all the figures $t = 0$ is the starting point when the representation is learned. Parameters: $n = 10, k = 3, \eta = 0.1, \sigma_1 = \sigma_2 = 0.01, T = 10^4$.

Results

Drift dynamics in linear Hebbian/anti-Hebbian networks

We first study drift in linear Hebbian/anti-Hebbian networks, which compress inputs into a lower dimensional space (Pehlevan, Hu, and Chklovskii 2015). While the resulting RFs are not localized, it is still instructive to study how learned representations evolve with noisy synaptic updates in this analytically tractable model.

The network we consider minimizes a similarity matching cost function (Pehlevan, Hu, and Chklovskii 2015). Here, the similarity between two vectors is defined as their dot product. Let $\mathbf{x}_t \in \mathbb{R}^n, t = 1, \cdots, T$ be a set of network inputs (or sensory stimulus) and $\mathbf{y}_t \in \mathbb{R}^k$, $k < n$ be the corresponding outputs constituting a representation. Similarity matching minimizes the mismatch between the similarity of pairs of inputs and corresponding pairs of outputs

$$\min_{\forall t:\mathbf{y}_t} \frac{1}{T^2} \sum_{t=1}^{T} \sum_{t'=1}^{T} (\mathbf{x}_t^\top \mathbf{x}_{t'} - \mathbf{y}_t^\top \mathbf{y}_{t'})^2. \tag{1}$$

Optimal solutions to this problem are given by projecting the inputs to their principal subspace (Pehlevan, Hu, and Chklovskii 2015). However, there is a continuum of such projections each corresponding to a basis in the subspace. This degeneracy can be seen from the rotational symmetry of the similarity matching cost function, (1). For any set of $\mathbf{y}_t$, $\mathbf{R}\mathbf{y}_t$ has the same cost, where $\mathbf{R}$ is an orthogonal matrix.

Previous work showed that this cost function can be minimized by a neural network in an online manner, where each input $\mathbf{x}_t$ is presented sequentially and an output $\mathbf{y}_t$ is produced (Pehlevan, Hu, and Chklovskii 2015) (Materials and Methods) by running the following neural dynamics until convergence:

$$\dot{\mathbf{y}}_t = \mathbf{W}\mathbf{x}_t - \mathbf{M}\mathbf{y}_t. \tag{2}$$

Here, $\mathbf{W}$ holds the feedforward synaptic weights and $\mathbf{M}$ the lateral weights. We note that at the fixed point of the neural dynamics $\mathbf{y}_t = \mathbf{M}^{-1}\mathbf{W}\mathbf{x}_t = \mathbf{F}\mathbf{x}_t$, where we define a filter matrix, $\mathbf{F} \equiv \mathbf{M}^{-1}\mathbf{W}$, whose rows are neural filters. After each presentation of an input and convergence of the neural dynamics, the weights $\mathbf{W}$ and $\mathbf{M}$ are updated with a Hebbian and anti-Hebbian rule, respectively:

$$\Delta\mathbf{W} = \eta(\mathbf{y}_t\mathbf{x}_t - \mathbf{W}), \quad \Delta\mathbf{M} = \eta(\mathbf{y}_t\mathbf{y}_t^\top - \mathbf{M}). \tag{3}$$

The learning rule is local in the sense that synaptic update only depends on activities of presynaptic and postsynaptic neurons. The update of $\mathbf{M}$ is anti-Hebbian due to the negation in (2). As the number of samples increases, these weights converge to a configuration where neural filters learn an orthonormal basis for the principal space (Pehlevan, Hu, and Chklovskii 2015; Pehlevan, Sengupta, and Chklovskii 2018).

Here, we model biological noise during learning by introducing noise to the weight updates, and examine the consequences of this noise. The updates are

$$\Delta\mathbf{W}(t) = \eta(\mathbf{y}_t\mathbf{x}_t^\top - \mathbf{W}(t)) + \boldsymbol{\xi}^W, \quad \Delta\mathbf{M}(t) = \eta(\mathbf{y}_t\mathbf{y}_t^\top - \mathbf{M}(t)) + \boldsymbol{\xi}^M, \tag{4}$$

where the noise terms ξ_{ij}^W, ξ_{ij}^M are independent Gaussian noise, with the following statistics: $\langle\xi_{ij}^W(t)\rangle = \langle\xi_{ij}^M(t)\rangle = 0$ and $\langle(\xi_{ij}^W(t)\xi_{kl}^W(t')\rangle = \eta\sigma_1^2\delta_{ik}\delta_{jl}\delta(t-t')$, $\langle\xi_{ij}^M(t)\xi_{kl}^M(t')\rangle = \eta\sigma_2^2\delta_{ik}\delta_{jl}\delta(t-t')$.

As expected, the network learns the principle subspace and maintains a stable performance, as quantified by the principle subspace projection (PSP) error: $||\mathbf{F}_t^\top\mathbf{F}_t - \mathbf{U}\mathbf{U}^\top||_F/||\mathbf{U}\mathbf{U}^\top||_F$ where $\mathbf{U}$ is a $n \times k$ matrix whose columns are the top k left singular vector of $\mathbf{X}$ (Fig. 2A). Due to the synaptic noise, network weights do not settle down to fixed points but roam around in the subspace that gives equally good solutions to the similarity matching problem. Consequently, the representation of a given stimulus $\mathbf{y}_t$ drifts over time (Fig. 2B). However, the similarity between any two outputs $\mathbf{y}_t$ and $\mathbf{y}_{t'}$ remains stable (Fig. 2C, SI Appendix Fig. S1A). As a

consequence, the drift does not change the length of the output vectors, $\mathbf{y}_t$, which undergo a random walk on a spherical surface (Fig. 2D). The drift of the representation ensemble $\mathbf{Y}_t \equiv [\mathbf{y}_1, \cdots, \mathbf{y}_T]$ behaves like a randomly-rotating rigid body consisting of a cloud of points (SI Appendix Fig. S1B).

These observations motivated us to quantify the drift rate by the rotational diffusion constant D_φ (Kämmerer, Kob, and Schilling 1997; Mazza et al. 2006) (Materials and Methods). We can derive an approximate analytical formula for D_φ from a linear stability analysis of the filter matrix ($\mathbf{F}$) (Materials and Methods, and SI Appendix for details):

$$D_\varphi \approx \frac{1}{8}\eta(\sigma_1^2 + \sigma_2^2)\sum_{i=1}^{k}\frac{1}{\lambda_i^2}. \tag{5}$$

Here, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ are the ordered eigenvalues of the input covariance matrix. (5) indicates that D_φ is proportional to the noise amplitude (Fig. 2F). Further, the drift amplitude along each eigenvector is proportional to $1/\lambda_i^2$. This is analogous to the rotation of an ellipsoid rigid body due to torque. In that system, the rotation around the axis with smaller moment of inertia is easier. Predictions of (5) is well in agreement with simulations (Fig. 2E-G).

The above simulation and analysis demonstrate that, in this model, while the network's output is drifting over time, similarity of representation is preserved. This is due to a coordinated random walk in the representational space, which in the linear case can be described by a rigid body rotation. This coordinated drift explores equally optimal representations. Because the quality of a representation is quantified by its representational similarity in (1), representational similarity is preserved despite the drift. Next, we consider nonlinear networks and show that these results carry over.

Drift dynamics in nonlinear Hebbian/anti-Hebbian networks

RFs of neurons in many brain areas are localized in the parameter space which they represent. For example, response of neurons in primary visual cortex (V1) are tuned to orientations of gratings (Hubel 1995). Neurons in the owl's external nucleus of the inferior colliculus (ICX) are tuned to different horizontal and vertical positions, forming an auditory spatial map (Peña and Konishi 2001). Place cells in hippocampus are active when an animal is at a particular spatial location of an environment (O'Keefe and Dostrovsky 1971).

A nonlinear version of our Hebbian/anti-Hebbian network can capture these localized RF properties (Fig. 1B). This network minimizes a nonnegative similarity matching (NSM) cost function (Földiak 1990; Pehlevan, Hu, and Chklovskii 2015; Sengupta et al. 2018):

$$\min_{\forall t:\mathbf{y}_t\geq 0}\frac{1}{T^2}\sum_{t,t'=1}^{T}(\mathbf{x}_t^\top\mathbf{x}_{t'} - \mathbf{y}_t^\top\mathbf{y}_{t'} - \alpha^2)^2, \tag{6}$$

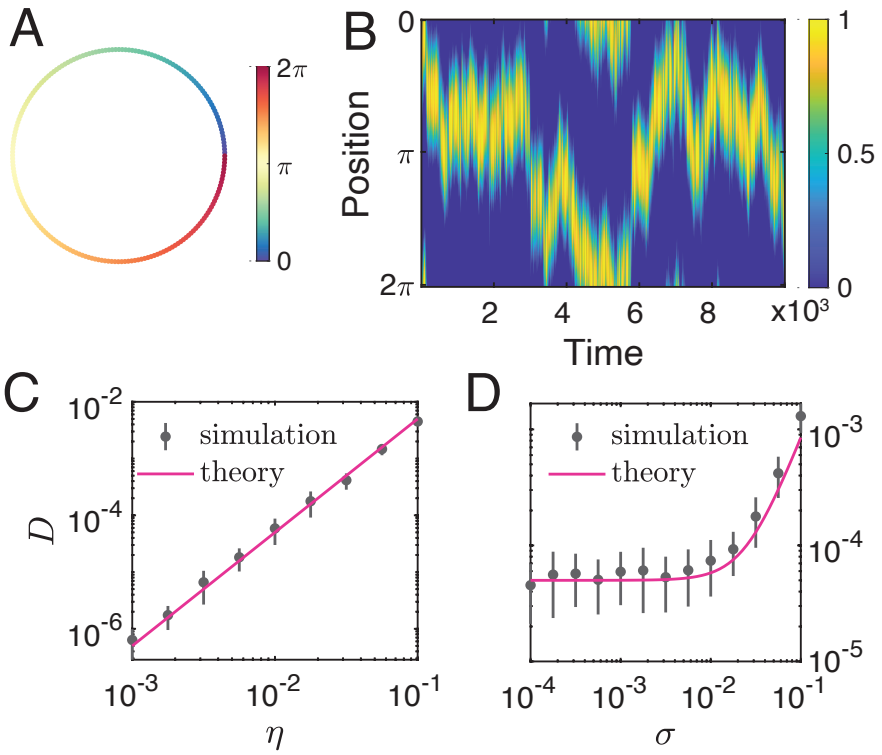

Figure 3: Drift of a single localized RF learned from a ring data manifold. (A) A ring in 2D as input dataset: $\mathbf{x}(\theta) = [\cos(\theta), \sin(\theta)]^\top, \theta \in [0, 2\pi)$. (B) The single RF has the shape of a truncated cosine curve, whose position drift on the ring and behaves like a random walk. (C,D) The effective diffusion constant D of centroid position increases with learning rate η even without explicit synatpic noise ($\sigma = 0$), and with the noise amplitude of explicit synaptic noise. Error bars represent standard deviation of 40 simulations. Megenta lines correspond to (9). Parameters: $\alpha^2 = 0, \beta_1 = \beta_2 = 0$, in (D) $\eta = 0.05$.

where $\mathbf{x}_t \in \mathbb{R}^n$ and $\mathbf{y}_t \in \mathbb{R}^k$ are input and output, respectively, and α^2 sets the threshold of similarity to be preserved in the output representation. With nonnegative neural activity, the above NSM objective function strives to preserve similarity for similar pairs of input samples but orthogonalizes the outputs corresponding to dissimilar input pairs. Compared with the linear network, non-negativity breaks the rotational symmetry of the solution, but the permutation symmetry is still preserved, i.e., exchanging identities of neurons does not change the objective function. To see this more clearly, the above target function can be written in terms of input-output Gram matrices: $||\mathbf{X}^\top\mathbf{X} - \mathbf{Y}^\top\mathbf{Y} - \alpha^2\mathbf{E}||_F^2$, where $\mathbf{X} \in \mathbb{R}^{n\times T}$, $\mathbf{Y} \in \mathbb{R}^{k\times T}$, and $\mathbf{E} \in \mathbb{R}^{T\times T}$ is a matrix with all entries set to 1. Thus, if $\mathbf{Y}$ is a solution, then $\mathbf{PY}$ is also a solution for any permutation matrix $\mathbf{P}$. Note that a general rotation would not preserve the nonnegativity of the output and thus not lead to a new solution. One can further control the behavior of learned representations by introducing regularizers to $\mathbf{y}_t$ in (6), for example an l_1 norm of $\mathbf{y}$ leads to more sparse code (Materials and Methods).

Similar to the previous linear Hebbian/anti-Hebbian network, this network also operates in an online fashion with a similar local learning rule. It takes an input $\mathbf{x}_t$ and generates an output $\mathbf{y}_t$ by running the following neural dynamics until it converges (Pehlevan 2019):

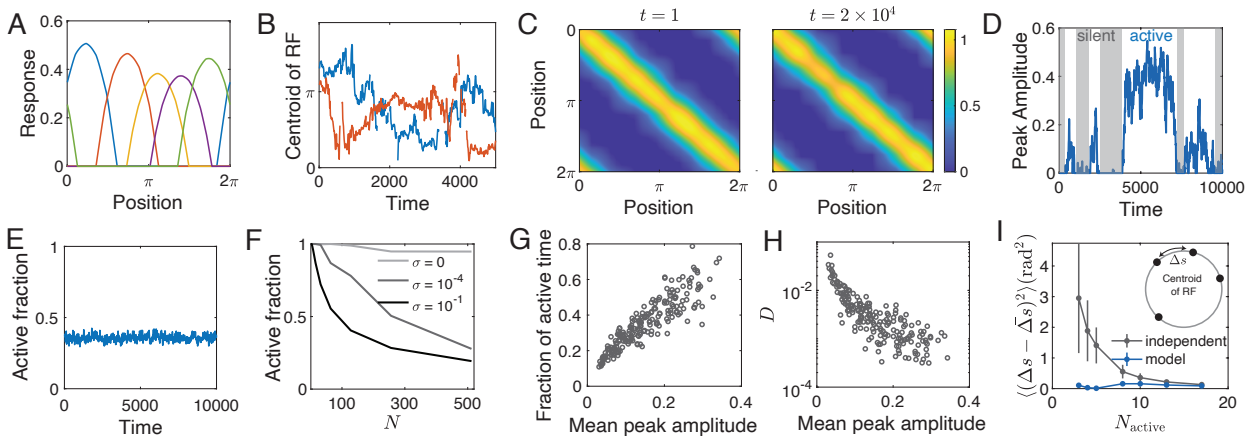

Figure 4: Drift of manifold-tiling localized RFs. (A) Learned localized RFs tile the input ring data manifold. Colors represent RFs of 5 example neurons. (B) Evolution of the RF centroids of two example neurons due to synaptic noise. (C) The representational similarity matrix $\mathbf{Y}^\top\mathbf{Y}$ is approximately circulant and stable over time. (D) When there are large number of neurons, each neuron has active and silent (shaded region) periods. (E) At population level, the fraction of neurons with active RFs are constant. (F) The fraction of neurons that have active RFs decreases with the total number of output neurons, as well as the noise amplitude. (G-H) Neurons that have stronger RFs tend to have longer active time (G) and also are more stable as quantified by smaller effective diffusion constants D (H). (I) At population level, the drift of RFs are coordinated such that their centroids are more uniformly distributed compared to that of the independent random walk case, in which the step size follows the same distribution of the Hebbian/anti-Hebbian network model. Shown are the variance of distances between adjacent centroids. Parameters in C-G: $N = 200, \sigma = 0.002, \eta = 0.05, \alpha^2 = 0$ except $\sigma = 0$ in B. In H: $\eta = 0.05$.

$$\frac{du_i}{d\tau} = -u_i + [\mathbf{W}\mathbf{x}_t]_i - \alpha b_i - [\bar{\mathbf{M}}\mathbf{y}_t]_i, \tag{7}$$
$$y_i = \max\{u_i/\mathrm{M}_{ii}, 0\},$$

where u_i and y_i represent the membrane potential and firing rate of neuron i, and b_i is the bias term. The forward weight matrix $\mathbf{W} \in \mathbb{R}^{k\times n}$ and lateral weight matrix $\mathbf{M} \in \mathbb{R}^{k\times k}$ (we have defined $\bar{\mathbf{M}} = \mathbf{M} - \mathrm{diag}(\mathbf{M})$) update according to the following "noisy" learning rule (Pehlevan 2019):

$$\Delta\mathbf{W} = \eta(\mathbf{y}_t\mathbf{x}_t^\top - \mathbf{W}) + \boldsymbol{\xi}^W, \quad \Delta\mathbf{M} = \eta(\mathbf{y}_t\mathbf{y}_t^\top - \mathbf{M}) + \boldsymbol{\xi}^M, \quad \Delta\mathbf{b} = \eta(\alpha\mathbf{y}_t - \mathbf{b}), \tag{8}$$

where η is the learning rate, and $\boldsymbol{\xi}^W$ and $\boldsymbol{\xi}^M$ are Gaussian white noise terms with the same statistics as in (4). The properties of the above learning rule without noise has been studied previously (Földiak 1990; Pehlevan 2019; Pehlevan and Chklovskii 2014; Pehlevan, Mohan, and Chklovskii 2017; Sengupta et al. 2018). Here, our interest is investigating how the learned representations evolve in the presence of noise in synaptic updates. We will first study the general drifting dynamics of RFs by a simple input: a ring data manifold. Based on the insights gained from this model, we will build models of drifting representations in place cells and neurons in PPC.

Localized receptive fields on a ring stimulus manifold

To explore how RFs evolve in the presence of synaptic noise, we first consider stimuli living on a ring (Fig. 3A), like the direction of a drifting grating used in experimental study of visual systems. Location of the stimulus on the ring is parameterized by an angular variable $\theta \in [0, 2\pi)$ (Materials and Methods). The input similarity matrix $\mathbf{X}^\top \mathbf{X}$ has a diagonal band structure and two inputs that are close on the ring have large similarity.

In the case of a single output neuron and in the absence of noise, the learned RF can be solved analytically, which is a truncated cosine curve centered around a random angle ϕ, i.e., $y_\phi(\theta) = \mu[\cos(\theta - \phi)]_+$ with μ being the peak amplitude. Derivation of this results and dependence of μ to model parameters is given in Materials and Methods. With synaptic noise during learning, the centroid drifts on the ring like a random walk. We quantified the speed of drift with an effective diffusion constant, D, of the centroid by the conventional relation: $\langle(\phi(t + \Delta t) - \phi(t))^2\rangle = 2D\Delta t$, where $\phi(t)$ is the centroid position at time t. For a single neuron, the dependence of D on the learning rate η and noise amplitude σ^2 can be analytically approximated as (Materials and Methods, and SI Appendix):

$$D \approx \frac{1}{2}(\eta^2 + 16\eta\sigma^2). \tag{9}$$

The first term of (9) is due to the sampling noise, i.e., the fact that the network sees one random stimulus at a time, and the second term is due to the explicit synaptic noise (noise terms in (8)). (9) indicates that faster learning and larger explicit noise result in more rapid drift of the RF. Numerical simulation agrees well with the theory (Fig. 3C-D).

When there is a population of output neurons, the Hebbian/anti-Hebbian network learns multiple localized RFs that tile the ring manifold with overlap (Fig. 4A), consistent with previous analytical accounts of simplified versions of such networks (Sengupta et al. 2018). With synaptic noise, we expect each RF to drift by a similar diffusion process as in the single neuron case, but with interactions between the neurons affecting the dynamics (Fig. 4B). In particular, the structure of neural population activity, as indicated by output similarity matrix $\mathbf{Y}^\top \mathbf{Y}$, is stable across time (Fig. 4C). Further, a neuron's response to the same stimulus is intermittent, having active and silent periods (Fig. 4D). At the population level, the fraction of neurons that have active RFs at any given time is constant (Fig. 4E), and it decreases with total number of output neurons as well as the noise amplitude (Fig. 4F). Thus, in a large population of neurons, only a small fraction of them will be active at a give time, forming a sparse population code. Neurons whose RFs have stronger tuning (characterized by the peak amplitude of RF) tend to be active more often (Fig. 4G), and have smaller drift (Fig. 4H).

At the population level, the drift of RFs are coordinated. To see the difference between our model and independent random walkers, we simulated N_{active} centroids undergoing indepen-

dent random walks on the ring. The step size of the independent random walks were drawn from the same distribution as the centroid shift between two adjacent time steps in our model (Materials and Methods). We observed that centroids in our model tile the ring manifold more uniformly than those of independent random walks, as indicated by the smaller variance of distances between two adjacent centroids on the ring (Fig. 4I).

Having gained better understanding of drifting dynamics in the above simple model, we now discuss models of representational drift in the Hippocampus CA1 region and PPC. The observations made in Fig. 4 will conceptually carry over, providing explanations for previous experimental observations.

A Hebbian/anti-Hebbian Network model for drifting place fields in CA1

CA1 place cells in the hippocampus play a crucial role in spatial memory and navigation. Recent long-term recording experiments show that place coding by the population of CA1 pyramidal cells are dynamic even when the animal is in the same familiar environment. In the time course of several weeks, some neurons lose their place fields while other previously non-place coding cells gain place fields. Despite the drift, the spatial information is preserved (Gonzalez et al. 2019; Y. Ziv et al. 2013).

One possible mechanism of place field formation is that CA1 place cells receive both forward input from grid cells in the entorhinal cortex and lateral competition from other place cells within the hippocampus (M.-B. Moser, Rowland, and E. I. Moser 2015). This motivated us to use the Hebbian/anti-Hebbian network to learn a place cell representation of a 2D square environment. Each position on the plane is represented by a population of grid cells with different grid spacing, phases and offsets (Fig. 5A, Materials and Methods), which serves as the input $\mathbf{x}_t$ to the network. After learning, some output neurons develop localized RFs (or place fields, Fig. 5B). This can be visualized by arranging each row of response matrix $\mathbf{Y}$ into a square matrix, as shown in Fig. 5B. The population of output neurons tile the 2D environment, as indicated by the uniform distribution of centroids of place fields (Fig. 5B, right). Due to the noise in the weight update, place fields continuously drift over time (Fig. 5C). Despite the drift, representational similarity of positions in the 2D environment is stable (SI Appendix, Fig. S2A). We also observed that a place cell may lose its place field for some time and gain a new place field later on (SI Appendix, Fig. S2B). The intermittence of RFs is due to both the competition between RFs and the fluctuation of $\mathbf{W}$ and $\mathbf{M}$. For example, once the forward input is smaller than the lateral inhibition at the centroid of an RF, then this RF becomes silent. The interval of these silent periods follow exponential distributions (SI Appendix Fig. S2C), indicating that the transition between active and silent state of RFs are random and memoryless. However, the fraction of neurons with active place fields at any given time remains constant (SI Appendix,

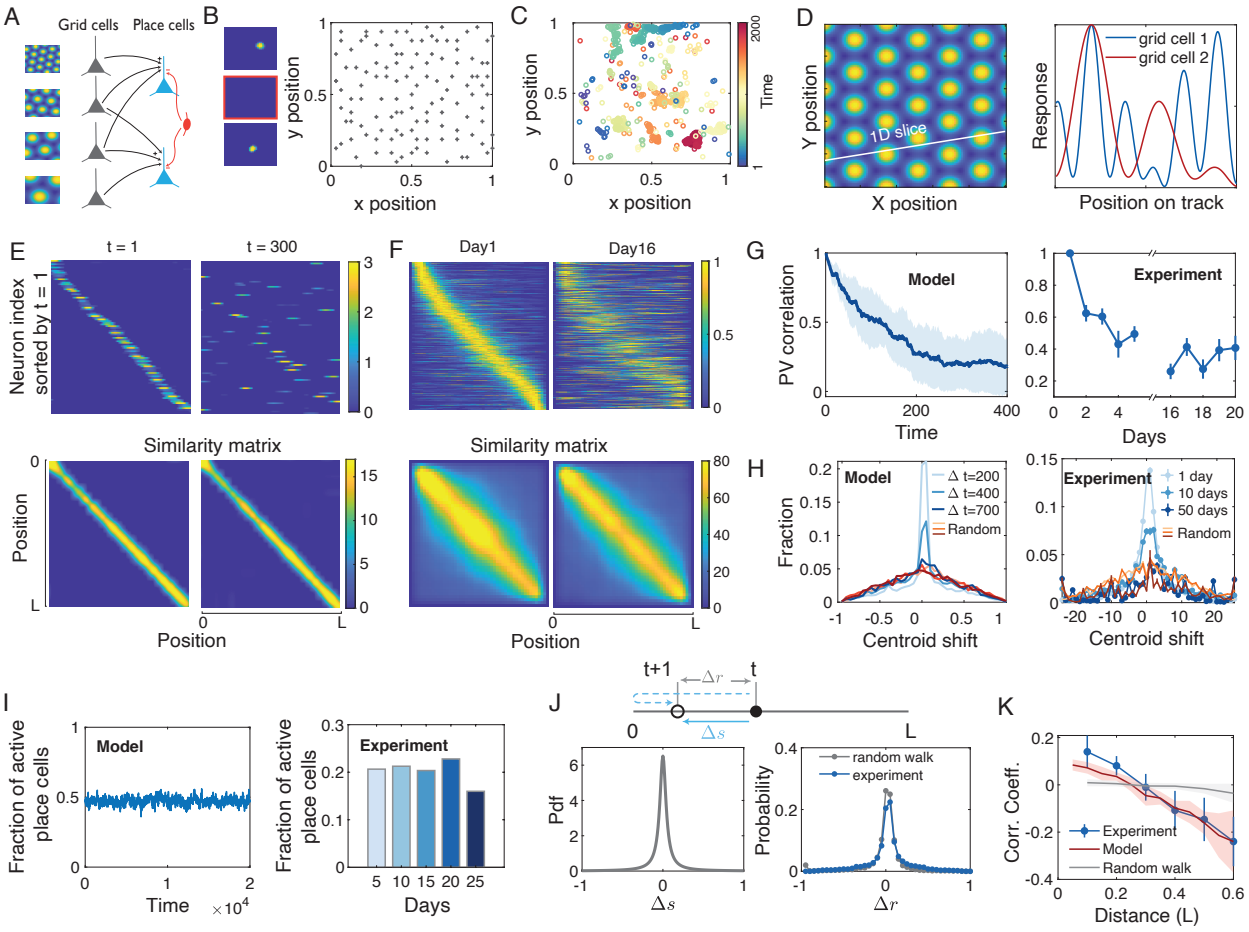

Figure 5: Drift of place fields. (A) Place cells receive input from grid cells that have different grid spacing, orientations and offsets. They also receive lateral inhibition due to competition with other place cells. (B) Left: learned place fields (left) tile the entitle 2D plane with red square highlighting an silent neuron. Right: each dot represents a centroid of a place field. (C) Drift of place field of an exemplar place cell. Each circle represents the position of its centroid at different times. (D) Left: Slice through a 2D grid field. Right: Response of neurons across this slice. (E) Upper: learned place fields tile a 1D linear track when sorted by their centroid positions (left), but continuously change over time (right). Lower: Representational similarity matrix $\mathbf{Y}^\top\mathbf{Y}$ of position is stable over time. (F) Experimental results corresponding to (E), place fields of a group of CA1 place cells concatenated from several mice when exploring the same familiar 1D linear track. (G) The average autocorrelation coefficient of population vectors representing each spatial position in the model (left) and experiment (right) decay over time. Shades represents standard deviation over different positions. (H) Probability distribution of centroid drifts of place cells at three different time intervals. Red lines represent random distributions, which are obtained by randomly aligning place fields of neurons between the same interval. The qualitative behavior of the model (left) is very similar to that of the experimental result (right). (I) Despite the continuous reconfiguration of place cell ensembles, the fraction of active place fields are stable over time in the model (left) and experiment (right). (J) Centroid shift $\Delta r = r(t+1) - r(t)$ observed in experiment could be a result of two different 'random walks' (blue lines) under reflecting boundary conditions (upper of J). To make a fair comparison with an independent random walk, we sample step sizes Δs of a random walk from a distribution $p(\Delta s)$ (lower left of F) that produces similar shift distribution as in experiment $p(\Delta r)$ (lower right panel of F). (K) Drifts of RFs show distance-dependent correlations, quantified by the average Pearson correlation coefficients. The model can recapitulate the behavior observed in experiment. Shades represent the standard deviation of different pairs of centroids. Experimental results in F-K are plotted using data from (Gonzalez et al. 2019). Parameters: (A)-(C) $\beta_1 = 0.1, \beta_2 = 0.05, \eta = 0.02, \sigma = 0.01, N_p = 400$. (D)-(I): $N_p = 200, \alpha = 15, \sigma = 0, \eta = 0.01, \beta_1 = 0, \beta_2 = 0.05$.

Fig. S2D).

As in the simple ring model in the previous section, we found that the drifts of individual RFs can be largely captured by a random walk but with intermittence due to the inactive periods of RFs. The drift speed can be quantified by an effective diffusion constant D. The dependence

of D on the number of neuron N is similar to the ring model in the previous section. When N is small, sampling noise dominates the synaptic update, resulting in a slight decrease of D as N increases. However, beyond a certain N, D increases rapidly with N (SI Appendix, Fig. S2E). Our model also predicts that neurons whose RFs have stronger tuning (larger peak amplitude of the RF) tend to be active more often, and have smaller drift (SI Appendix,Fig. S2F,G).

While the above predictions could be compared with long-term recording experiments for animals in a 2D environment, existing long-term recording experiments are limited to 1D environments (typically linear tracks). To compare our model with these experimental results, we simulated our model in a 1D environment, where grid cell responses are modeled as 1D slices through the 2D grid fields, as observed in experiments (Yoon et al. 2016)(Fig. 5D, Material and Methods). The model generates qualitatively similar results as the above 2D place cell model. The learned place fields tile the linear track but drift over time due to ongoing noisy weight updates, yet the representational similarity is stable over time (Fig. 5E). This is also observed in an experiment (Gonzalez et al. 2019), where CA1 pyramidal cells were recorded when mice were in the same familiar environment for several months (Fig. 5F). Due to drifting place fields, the autocorrelation coefficients of neural population vectors in both our model and in experiment decay over time (Fig. 5G). The shift of centroids of place fields increases with time, with a distribution eventually approaching the case wherein the place fields are randomly permuted. Such behavior closely resembles that of experiments (Gonzalez et al. 2019) (Fig. 5H). Despite the continuous reconfiguration of the neural assemblies representing the position, the fraction of active place cells is stable over time (Fig. 5I).

To further explore the underlying structure of centroid shifts, and test the main prediction of our model that the drift of RFs is coordinated, we set out to compare the experiment and our simulation results to a null hypothesis — the shift of RFs behave like an independent random walk. To make a fair comparison, for the null hypothesis, we assume each centroid takes a step size Δs that is drawn from a distribution $p(\Delta s)$ with a reflecting boundary condition (upper panel of Fig. 5J). The distribution $p(\Delta s)$ was chosen such that the resulting centroid shift Δr closely matches that of experiment (lower left panel of Fig. 5J, Material and Methods). Centroid shifts in experiment show clear distance-dependent correlations, i.e., two RFs that are very close to each other are more likely to drift in the same direction on the next day, while RFs that are far apart are more likely to drift in opposite directions (blue line, Fig. 5 K). This is in stark contrast with the independent random walk picture (gray horizontal line, Fig. 5 K), but can be recapitulated by our model (red line, Fig. 5 K), suggesting that the drift of RFs in experiment is coordinated at the population level, possibly to preserve representational similarity.

A Hebbian/anti-Hebbian network model for drifting RFs of neurons in PPC

The above model and results can be extended to another sensorimotor task, in which mice were trained to navigate a virtual T-maze (Driscoll et al. 2017). At the first half of the T-stem, mice saw one of two alternative visual scenes, and associated them with left-turn or right-turn at the T-junction to receive a reward at the end of the track (Fig. 6A). PPC is essential for this task (Driscoll et al. 2017; Harvey, Coen, and Tank 2012). After learning, a sub-population of neurons in PPC have localized receptive fields, i.e., they fire when a mouse is at a specific position along the T-maze and their RFs tile the T-maze, providing essential information about the task. While mice stably perform the task after learning, the neural population activities in PPC continuously drift over weeks. Despite such drift, the task information can be stably encoded by the activities of a subpopulation of PPC neurons (Driscoll et al. 2017).

To better understand the positional tuning and drifting behavior of neurons in PPC, we again use a Hebbian/anti-Hebbian network with noisy weight update rule to model this system. For simplicity, the task information is represented by a vector $\mathbf{x}_{R/L}(\theta) = [\cos(\theta), \sin(\theta), \pm 1]^\top$, $\theta \in [0, \pi)$, with the last entry indicating a right-turn (1) or a left-turn task (-1).

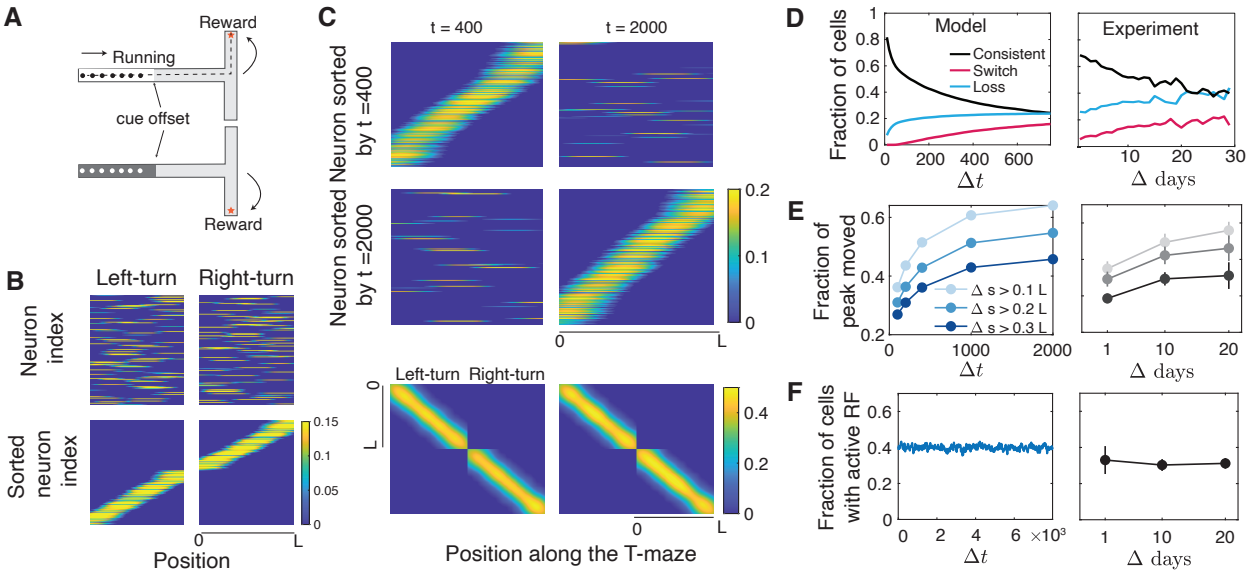

Figure 6: Representational drift in PPC. (A) Schematic of the visual-cue-guided T-maze sensorimotor task as in (Driscoll et al. 2017). The linear length of the track from the beginning to the end (dashed line) is L. (B) Population activity for the left-turn and right-turn task before (upper) and after (lower) sorting based on the centroids of their RFs. Only neurons that have active RFs at the given time point are shown. (C) Population activity drifts but representational similarity is stable over time. Activity of neurons identified with significant peak in the sorted time (upper and middle). Representational similarity matrix is stable for both left-turn and right-turn task (lower panels). (D) left: For a group of neurons that have tuning to left-turn (or right-turn) tasks, the fraction of them that have consistent tuning (black), switched tuning (magenta), losing tuning (cyan) to left (or right) in the following time. (E) Shift of RFs for neurons with a significant peak between time t and $t + \Delta t$. Smaller shift happens more often than larger shift. (F) The fraction of the neurons with active RFs is stable across time. In (D)-(F) Left panels are simulation results of our model, right panels are corresponding experiment results from (Driscoll et al. 2017). Parameters: $N = 400, \alpha^2 = 1.6, \eta = 0.05, \sigma = 10^{-4}, \beta_1 = 10^{-4}, \beta_2 = 10^{-3}$.

After learning, the population of output neurons in the model develop positional tuning of the T-maze, i.e., for either left-turn input $\mathbf{x}_L$ or right-turn input $\mathbf{x}_R$, there are a subpopulation of neurons that fire most strongly when the animal is at the specific positions of the track, forming RFs that tile the maze (Fig. 6B).

To see how the RFs of neurons evolve over time, we first sort neurons with significant RFs based on the centroid positions of their RFs at a reference time point. We find that RFs of neurons drift over time, i.e., neurons rarely have the same or similar RFs at two long-separated time points. However, the population representation of location-context information is stable across time. Thus, at any given time, we can identify a subset of neurons with significant RFs that tile the positions of the T-maze for both left-turn and right-turn tasks (Fig. 6C, upper and middle panels). Despite the drift, representational similarity of both left-turn and right-turn tasks are stable over time (Fig. 6C, lower panel). Neurons also gradually change their tuning to tasks choices. For example, a group of neurons that are tuned to the left-turn tasks at time 0 may loss such tuning or become tuned to right-turn tasks, and vise versa. Drift of an RF accumulates over time, such that the probability of centroid shift that is larger than a certain distance increases with time (left, Fig. 6E). Overall, the fraction of neurons that have positional tuning at any time for both left-turn and right-turn trials are constant (left, Fig. 6F). All these behavior are consistent with the experiment (right panels of Fig. 6D-F). Together, these comparisons shows that our simple model can explain many characteristics of representational drift in PPC.

Summary and Discussion

In this paper, we explored the hypothesis that representational drift is due to the existence of many (possibly infinite) ensembles of population codes that achieve a representational objective. Noise in learning drives the network to explore this space, causing the drift of population activity. While our focus was on synaptic noise, other sources of noise can also cause similar representational drift with potentially different statistics. Similarly, network architectures that optimize other objective functions can also show drift when learning with noise. However, we expect the drift to be strongly affected by the degeneracy of the solution space of the objective function. For example, in a feedforward network performing online principal component analysis, which has no degeneracy as the principal subspace projection task, we found stabilized representations in the presence of noise (Fig. S3, SI Appendix).

To explore the consequences of our hypothesis in a concrete model, we focused on a well-studied model for biologically plausible representation learning that optimizes similarity-based representational objectives (Pehlevan and Chklovskii 2019). We showed that simple

Hebbian/anti-Hebbian networks with noisy synaptic updates recapitulate observed representational drift phenomena in experiments. In the case that the network consists of a single output neuron, we observed that its RF behaves like a random walk on the data manifold with a diffusion constant that depends on the noise amplitude, learning rate, and statistical structure of the input. When the network consists of many neurons, different drifting RFs are coordinated such that representational similarity is stable across time.

We used the Hebbian/anti-Hebbian network as a simplified model to study the RFs of hippocampal place cells and neurons in the PPC. Our model recapitulates the drift statistics at population-level observed in these regions: First, a constant fraction of active neurons represent task variables at a given day. Second, neurons drop in and out of this assembly over days. As a consequence, the autocorrelation coefficient of population vectors decay over time. Third, drift at population level preserves representational similarity (Fig. 5,6). While simple, the network captures the essential properties of those neural circuits, i.e., RFs are shaped by input from upstream and effective lateral inhibition/competition within the layer. It is also possible to model these systems by training a general recurrent neural network (RNN), as has been demonstrated in (Rajan, Harvey, and Tank 2016). It will be interesting to see whether neurons in such RNN models with noisy weight update also show representational drift.

Our model makes several testable predictions. First, our model predicts that the drifts of RFs are coordinated. This coordination is arising from the existence of a representational objective for the neural population as a whole. We verified this prediction in hippocampal data Fig. 5J,K. Second, it predicts that neurons whose synapses have faster turnover dynamics tend to drift more rapidly. For example, the lifetime of spines of pyramidal cells in hippocampus is about 1 to 2 weeks, much shorter than that of neocortex neurons (Attardo, Fitzgerald, and Schnitzer 2015). This suggests that representational drift should be more prominent in hippocampus than in neocortex. Furthermore, the lifetime of synapses can be perturbed by blocking receptors such as NMDA (Zuo et al. 2005), which will alter the stability of RFs. A definitive examination of this prediction requires experiments that both measure the life time of synapses and the long-term neural activity in brain regions that represent learned stereotyped behavior under unperturbed and perturbed states. While challenging, this is nonetheless becoming within reach with new experimental techniques. Third, our model predicts that neurons with strongly tuned RFs should be more stable. This prediction can be tested by examining the amplitude of tuning curves (RFs) of individual neurons and their stability in long-term recording experiments. Furthermore, the strength of RFs can be perturbed by optogenetic tools to examine how it affects the stability of RFs.

Representational drift contradicts the hypothesis that stable neural activity is the substrate of stable behavior. However, there needs to be stable aspects of representations which provide a substrate for stable downstream decoding and readout. Representational similarity can

be one such substrate for multiple reasons. First, our modeling shows that achieving stable representational similarity despite the drift of population activity is biologically plausible. Second, stable representational similarity may be a general internal structure of drifting neural population activity. For example, mouse visual cortices show strong representational drift yet the relation between population activities that represent different inputs remains stable and stereotyped (Deitch, Rubin, and Y. Ziv 2021). Conserved and stable internal structure of neural activity has also been discovered in hippocampus and prefrontal cortex in free-behaving mice (Rubin et al. 2019). Third, experimental evidence is consistent with stable representational similarity being a foundation for robust downstream decoding. Studies in monkey motor cortices have shown that stable geometry of latent population dynamics underlies stereotyped reaching tasks (Gallego et al. 2020) despite the inherently variable single neuron activities (Liberti et al. 2016; Rokni et al. 2007) (see however (Chestek et al. 2007; Katlowitz, Picardo, and Long 2018)). Interestingly, a recent experiment has shown that the spatial code of different environments in the hippocampus are random in individual rodents but share the same geometry across different animals (Kinsky et al. 2018). Finally, preserving pairwise similarity of representations may provide some computational benefits. Recent unsupervised learning algorithms for image recognition, such as contrastive representational learning (Chen et al. 2020) and "Barlow Twins" (Zbontar et al. 2021), are based on objectives that maximize representational similarity between a sample and its distorted/augmented versions. Such algorithms can achieve comparable performance to supervised learning algorithms. From a theoretical point of view, the representational similarity matrix (or kernel) determines the number of sampled stimuli required to learn an accurate linear readout from a population code, indicating that performance need not suffer as long as the representational kernel is preserved (Bordelon and Pehlevan 2021).

A hypothesis for achieveing stable readout despite time-varying neural activity is that the variation happens in the "coding-null space" (Druckmann and Chklovskii 2012; Kaufman et al. 2014). Representations in our model exhibit drift in all dimensions, precluding the existence of such coding-null space. Similarly, a closer scrutiny of the response of PPC neurons in the 'T-maze' task showed that drift is not confined to a "coding null space" (Rule, Loback, et al. 2020). Hence, an adaptive readout mechanism which involves synaptic plasticity to track and compensate the drift is required to achieve stable behavior (Rule, Loback, et al. 2020; Rule and O'Leary 2021). Whether and how such a mechanism is implemented in the brain remains an open question.

The ubiquity of representational drift raises the question of whether it is a biological feature or a bug. Representational drift may be desirable under certain circumstances (Mau, Hasselmo, and Cai 2020). For example, in a model of the bird song learning system, variation in the neural representation of the stereotyped behavior enables the system to adapt quickly to

a shift of target song, and to reduce error due to loss of neurons (Duffy et al. 2019). Drift can accommodate new learning with minimal inference by continuously modifying existing memories (Mau, Hasselmo, and Cai 2020). Other authors proposed that noisy synaptic plasticity and spine motility enable cortical networks of neurons to carry out probabilistic inference by sampling from a posterior distribution of network configurations (Kappel et al. 2015). Such sampling would lead to a representational drift as a byproduct.

Material and Methods

Similarity matching and the linear Hebbian/anti-Hebbian network

The linear Hebbian/anti-Hebbian network can be derived from (1). The detailed derivation can be found in (Pehlevan, Hu, and Chklovskii 2015; Pehlevan, Sengupta, and Chklovskii 2018), we sketch the main steps here. Starting from the cross term in (1), by introducing a new matrix variable $\mathbf{W} \in \mathbb{R}^{k\times n}$, we obtain

$$-\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\mathbf{y}_t^\top \mathbf{y}_{t'}\mathbf{x}_t^\top \mathbf{x}_{t'} = -\frac{1}{T^2}\sum_{t=1}^{T}\mathbf{y}_t^\top\left[\sum_{t'}^{T}\mathbf{y}_{t'}\mathbf{x}_{t'}^\top\right]\mathbf{x}_t = \min_{\mathbf{W}\in\mathbb{R}^{k\times n}} -\frac{2}{T}\sum_{t=1}^{T}\mathbf{y}_t^\top\mathbf{W}\mathbf{x}_t + \mathrm{Tr}\mathbf{W}^\top\mathbf{W}. \tag{10}$$

Similarly, we can introduce another matrix variable $\mathbf{M}$ for the quartic $\mathbf{y}_t$ term in (1):

$$-\frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T}\mathbf{y}_t^\top \mathbf{y}_{t'}\mathbf{y}_t^\top \mathbf{y}_{t'} = \frac{1}{T^2}\sum_{t=1}^{T}\mathbf{y}_t^\top\left[\sum_{t'}^{T}\mathbf{y}_{t'}\mathbf{y}_{t'}^\top\right]\mathbf{y}_t = \max_{\mathbf{M}\in\mathbb{R}^{k\times k}} \frac{2}{T}\sum_{t=1}^{T}\mathbf{y}_t^\top\mathbf{W}\mathbf{y}_t - \mathrm{Tr}\mathbf{M}^\top\mathbf{M}. \tag{11}$$

By substituting (10) and (11) into (1) and changing orders of optimization (Pehlevan, Sengupta, and Chklovskii 2018) we get:

$$\min_{\mathbf{W}\in\mathbb{R}^{k\times n}}\max_{\mathbf{M}\in\mathbb{R}^{k\times k}}\frac{1}{T}\big[2\mathrm{Tr}(\mathbf{W}^\top\mathbf{W}) - \mathrm{Tr}(\mathbf{M}^\top\mathbf{M}) + \min_{\mathbf{y}_t\in\mathbf{R}^{k\times 1}} l_t(\mathbf{W},\mathbf{M},\mathbf{y_t})\big], \tag{12}$$

where

$$l_t(\mathbf{W},\mathbf{M},\mathbf{y_t}) = -4\mathbf{x}_t^\top\mathbf{W}\mathbf{y}_t + 2\mathbf{y}_t^\top\mathbf{M}\mathbf{y}_t. \tag{13}$$

The minimax problem (12) can be solved by the following two-step online algorithm. First, minimizing (13) while keeping $\mathbf{W}$ and $\mathbf{M}$ fixed, which is solved by running the dynamics of output variable $\mathbf{y}_t$ until convergence

$$\frac{d\mathbf{y}_t}{dt} = \mathbf{W}\mathbf{x}_t - \mathbf{M}\mathbf{y}_t. \tag{14}$$

Second, after the convergencec of $\mathbf{y}_t$, update $\mathbf{W}$ and $\mathbf{M}$ by gradient descent and gradient ascent of (12) respectively:

$$W_{ij} \leftarrow W_{ij} + \eta(y_i x_j - W_{ij}), \quad M_{ij} \leftarrow M_{ij} + \eta(M_{ij} - y_i y_j). \tag{15}$$

The above learning algorithm (14), (15) can be naturally mapped onto a single-layer biologically plausible neural network, the linear Hebbian/anti-Hebbian network. Here, $\mathbf{y}_t$ is the neural activity of the output, $\mathbf{W}$ and $\mathbf{M}$ are synaptic matrices of the forward and lateral connections respectively. The synaptic update rule (15) is local since the change of a synapse only depends on the activity of presynaptic and postsynaptic neurons.

Calculation of the rotational diffusion constant

An analytical calculation of the rotational diffusion constant, defined by (Hunter et al. 2011; Kämmerer, Kob, and Schilling 1997; Mazza et al. 2006),

$$D_\varphi \equiv \lim_{t\to\infty} \frac{1}{4t}\langle|\vec{\varphi}(t) - \vec{\varphi}(0)|^2\rangle. \tag{16}$$

is difficult. However, we were able to obtain an approximation which matches numerical experiments very well, as shown in Fig. 2E-G. We present the details of this derivation in the SI Appendix. Our approximation assumes that 1) angular displacements of the representation vectors after different time steps are not correlated, and 2) the network weights stay close to the optimal representation manifold. Under these assumptions, D_φ can be approximated by the mean squared angular displacement (MSAD),

$$D_\varphi \approx \frac{1}{4\Delta t}\langle|\Delta\vec{\varphi}|^2\rangle, \tag{17}$$

where Δt is the small time interval elapsed during a single step update, and $\Delta\vec{\varphi}$ arises from a noisy synaptic update to the network with an optimal set of synapses. We calculate MSAD analytically (SI) to arrive at (5).

To numerically estimate D_φ from trajectory of $\mathbf{y}(t)$ with total length of T time steps, we first calculate $\delta\vec{\varphi}$ at each simulation step, then estimate $\vec{\varphi}(t)$ by cumulatively summing $\delta\vec{\varphi}$ up to time step t. Next, we estimate the MSAD of time interval τ using all the pairs of $\vec{\varphi}(t+\tau)$ and $\vec{\varphi}(t)$, which gives $\langle|\Delta\vec{\varphi}|^2\rangle = \langle|\vec{\varphi}(t+\tau) - \vec{\varphi}(t)|^2\rangle$. Last, we plot $|\Delta\vec{\varphi}|^2$ as a function of τ and fit a line that pass the origin to the data. The slope of the best fit is then $4D_\varphi$.

Hebbian/anti-Hebbian network and nonnegative similarity matching

The nonlinear Hebbian/anti-Hebbian network (Eq. s(7) and (8)) can be derived from the general nonnegative similarity matching (NSM) problem (Pehlevan 2019; Pehlevan and Chklovskii 2019). Denoting the input data as a set of vectors $\mathbf{x}_{t=1,\cdots,T} \in \mathbb{R}^n$ and the corresponding output vectors $\mathbf{y}_{t=1,\cdots,T} \in \mathbb{R}^k$, the NSM objective is defined as

$$\min_{\forall \mathbf{y}_t \geq 0} \frac{1}{T^2}\sum_{t,t'=1}^{T}(\mathbf{x}_t^\top \mathbf{x}_{t'} - \mathbf{y}_t^\top \mathbf{y}_{t'} - \alpha^2)^2 + \frac{1}{T}\sum_{t=1}^{T}(2\beta_1||\mathbf{y}_t||_1 + \beta_2||\mathbf{y}_t||_2^2), \tag{18}$$

where α^2 sets the threshold of similarity to be preserved in the output representation, the other two regularizers β_1, β_2 control the sparsity and amplitude of output. The detailed derivation of (7) and (8) from (18) is described in (Pehlevan 2019).

To see why the above NSM objective (18) leads to localized RFs, we can consider the simpler case where $\beta_1 = \beta_2 = 0$ and a single pair of inputs. If two inputs are similar, i.e., $\mathbf{x}_1 \cdot \mathbf{x}_2 > \alpha^2$, then the corresponding outputs $\mathbf{y}_1$ and $\mathbf{y}_2$ would prefer $\mathbf{y}_1 \cdot \mathbf{y}_2 = \mathbf{x}_1 \cdot \mathbf{x}_2 - \alpha^2$, i.e., they are also similar. In contrast, if two inputs are less similar, i.e., $\mathbf{x}_1 \cdot \mathbf{x}_2 < \alpha^2$, due to the nonnegativity of outputs, $\mathbf{y}_1$, $\mathbf{y}_2$ they tend to be orthogonal: $\mathbf{y}_1 \cdot \mathbf{y}_2 = 0$. To achieve this, dissimilar inputs must activate non-ovelapping sets of neurons. Thus, in manifold learning, (18) preserve local geometric structure in the $\mathbf{y}$ representation space of the input data clouds. A detailed explanation of why localized RFs are learned in a simplified version of (18) is provided in (Sengupta et al. 2018).

The neural dynamics derived from (18) (with regularizers) differ from that in main text slightly by changing the transfer function in (7) to

$$y_i = \max\{(u_i - \beta_1)/(\beta_2 + \mathrm{M}_{ii}), 0\}. \tag{19}$$

Derivation of the diffusion constant of the ring model

We sketch the derivation of (9) here, more details are in SI Appendix. We again consider the approximation that the diffusion coefficient can be approximated by the mean squared displacement around a fixed point by a noisy synaptic update.

Consider a single output neuron that learns a RF from inputs that are on a ring manifold (Fig. 3A). The response of the output neuron to $\mathbf{x} = [\cos\theta, \sin\theta]^\top$ is

$$y(\theta) = \frac{1}{m+\beta}[w_1 \cos\theta + w_2 \sin\theta - \alpha b]_+, \tag{20}$$

where $[x]_+$ denotes the rectified linear function and β is the l_2 regularizer. The stationary state parameters $\{w_1^*, w_2^*, m^*, b^*\}$ satisfy the following conditions

$$\begin{aligned} \langle w_1^* \rangle &= \langle y(\theta)\cos\theta \rangle_\theta, \quad \langle w_2^* \rangle = \langle y(\theta)\sin\theta \rangle_\theta, \\ \langle m^* \rangle &= \langle y^2(\theta) \rangle_\theta, \quad \langle b^* \rangle = \alpha \langle y(\theta) \rangle_\theta. \end{aligned} \tag{21}$$

These equations can be solved self-consistently by assuming an ansatz of the form:

$$y_\phi(\theta) = \mu[\cos(\theta - \phi) - \cos(\psi)]_+, \quad \theta \in (-\pi, \pi], \tag{22}$$

which gives the dependence of μ and ψ on α, β

$$\mu^2 = \frac{2\psi - \sin 2\psi - 4\beta\pi}{4\psi + 2\psi\cos 2\psi - 3\sin 2\psi}, \quad \alpha^2 = \frac{\cos\psi(2\psi - \sin 2\psi)}{4(\sin\psi - \psi\cos\psi)}. \tag{23}$$

Using the fact that $dy(\theta)/d\theta = 0$ at $\theta = \phi$, we have $\tan\phi = w_2^*/w_1^*$ and

$$\frac{d\phi}{dt} = \frac{1}{\hat{\mu}}\left(\frac{dw_2}{dt}\cos\phi - \frac{dw_1}{dt}\sin\phi\right), \tag{24}$$

where $\hat{\mu} = \sqrt{w_1^{*2} + w_2^{*2}}$ is the norm of weight vector. Using the noisy update rule (8) and (22), the shift of centroid due to one-step update eventually becomes

$$\Delta\phi = \frac{1}{\hat{\mu}}\{\eta\mu[\cos(\theta-\phi) - \cos\psi]_+ \sin(\theta-\phi) + (\xi_2\cos\phi - \xi_1\sin\phi)\}. \tag{25}$$

Finally, using the relation $\langle(\Delta\phi)^2\rangle \approx 2D\Delta t$, we have

$$D \approx \gamma\eta^2 + \frac{\eta\hat{\sigma}^2}{\hat{\mu}^2}, \tag{26}$$

where $\hat{\sigma}^2 \equiv \sigma_1^2\cos^2\phi + \sigma_2^2\sin^2\phi$, and

$$\gamma \equiv \frac{\mu^2}{\hat{\mu}^2}\langle([\cos(\theta-\phi) - \cos\psi]_+^2 \sin^2(\theta-\phi))\rangle_\theta. \tag{27}$$

When $\alpha = \beta = 0, \sigma_1 = \sigma_2 = \sigma$, we have $\gamma = 1$ and $\hat{\mu} = 1/4$, (26) reduces to (9) in the main text.

Numerical simulation of 2D place cells

We considered a 32×32 grid plane as the environment, each position (x, y) is represented by a group of grid cells with different grid spacings, orientations and offsets as observed in experiment (Stensola et al. 2012). The hexagonal firing fields of grid cells are modeled as a summation of three two-dimensional sinusoidal functions as in (Kropff and Treves 2008; Lian and Burkitt 2020; Solstad, E. I. Moser, and Einevoll 2006)

$$G(\boldsymbol{r}) = \frac{2}{3}\left(\frac{1}{3}\sum_{i=1}^{3}\cos\left(\frac{4\pi}{\sqrt{3}l}\boldsymbol{e}_i\cdot(\boldsymbol{r}-\boldsymbol{r_0})\right) + \frac{1}{2}\right), \tag{28}$$

where $\boldsymbol{r} = [x, y]^\top$ is the location on the plane, $\boldsymbol{r}_0 = [x_0, y_0]^\top$ is the phase offset, l is the grid spacing. $\boldsymbol{e}_i = (\cos(2\pi i/3 + \theta), \sin(2\pi i/3 + \theta)), i = 1, 2, 3$ is the unit vector in the direction $2\pi i/3 + \theta$ with θ being the grid orientation. In the simulation, grid cells have 5 modules, i.e., $N_l = 5$. The value of l increases as geometric series with a ratio 1.42 that is consistent with experiment (Stensola et al. 2012). For example, in our simulation, the smallest spacing is $0.2L$ with L being the linear length of the plane, then the rest spacing would be $0.2 \times 1.42L, \cdots, 0.2 \times 1.42^{N_l-1}L$. In each module, the number of orientation θ, $N_\theta = 6$, which are drawn uniformly in the range $[0, \pi/3)$. Similarly, the number of grid phase offsets x_0, y_0 are $N_x = 5$ and $N_y = 5$, which are drawn uniformly in the range $[0, l)$. As result, the total number of grid cell is $N_g = N_l N_\theta N_x N_y = 750$.

Numerical simulation of 1D place cell

We consider a linear track with length L. Tuning curves of grid cells on the linear track are slices through 2D grid fields described above. The orientation of the slices are the same and randomly selected in the range $[0, \pi/3]$.

Autocorrelation coefficient of the population vector

In all the cases, the autocorrelation coefficient ρ of population vector is defined as the Pearson's correlation coefficent between $\mathbf{y}_t$ and $\mathbf{y}_0$ to the same input:

$$\rho(t) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{y_{0,i} - \bar{y_{0,i}}}{\sigma_{y,0}} \right) \left(\frac{y_{t,i} - \bar{y_{t,i}}}{\sigma_{y,t}} \right), \tag{29}$$

where $\bar{y_0}, \bar{y_t}$ are the mean of $y_{0,i}$ and $y_{t,i}$, $\sigma_{y,0}, \sigma_{y,t}$ are the standard deviation of $y_{0,i}$ and $y_{t,i}$.

Step size in independent random walks place fields

In Fig. 5J,K, the step size of independent random walks were drawn from a distribution $p(\Delta s)$ closely matching that of experiment. To determine this distribution, we first calculated the distribution of centroid shift between two adjacent days in experiment $p(\Delta r)$ with $\Delta r = r(t+1) - r(t)$. For a random walk whose centroid is at position $\hat{r}_t$, its position at next time step is $\hat{r}_{t+1} = \hat{r}_t + \Delta s$ with Δs randomly sampled from $p(\Delta s)$. To constrain $\hat{r}_{t+1}$ in the range of the track $[0, L]$ with L being the length of the track, we assumed a reflecting boundary condition, which gives

$$\hat{r}_{t+1} = \begin{cases} |\hat{r}_t + \Delta s| & \hat{r}_t + \Delta s < 0 \\ 2L - (\hat{r}_t + \Delta s) & \hat{r}_t + \Delta s > L \\ \hat{r}_t + \Delta s & \text{other} \end{cases}$$

The shift of centroid in the random walk model is then determined by $\Delta\hat{r} = \hat{r}_{t+1} - \hat{r}_t$ according to the above equation. Our aim is to find a distribution $p(\Delta s)$, such that $p(\Delta\hat{r})$ is close to that of experiment $p(r)$. Based on the shape of $p(r)$, we searched $p(\Delta s)$ from a family of Levy's alpha stable distribution (Samorodnitsky and Taqqu 2017) by minimizing the Kullback–Leibler divergence between $p(r)$ and $p(\hat{r})$.

Data source and processing

Experimental data presented in Fig. 5 are originally described in (Gonzalez et al. 2019). We used the processed data and MATLAB code, which are available at the Caltech Research Data Repository (https://doi.org/10.22002/d1.1229) to produce these plots.

Experimental data presented in Fig. 6 is extracted from Figure 2C, 2D, and 4E of (Driscoll et al. 2017). The data is freely available in (Driscoll et al. 2020).

Acknowledgements

This work was supported by NIH (1UF1NS111697-01), the Intel Corporation through Intel Neuromorphic Research Community, and a Google Faculty Research Award. We thank Walter Gonzalez, Hanwen Zhang, Anna Harutyunyan and Carlos Lois for sharing the data on place cell recordings. We thank Laura Driscoll, Noah Pettit, Matthias Minderer, Selmaan Chettih and Christopher Harvey for making the T-maze experimental data available. We are grateful to members of Pehlevan group for helpful discussions, and Christopher Harvey for comments on the manuscript.

Competing interests

The authors declare no competing interests.

References

Atick, J. J. and A. N. Redlich (Mar. 1992). "What Does the Retina Know about Natural Scenes?" In: *Neural Computation* 4.2, pages 196–210. DOI: 10.1162/neco.1992.4.2.196.

Attardo, Alessio, James E Fitzgerald, and Mark J Schnitzer (2015). "Impermanence of dendritic spines in live adult CA1 hippocampus". In: *Nature* 523.7562, pages 592–596.

Attneave, Fred (1954). "Some informational aspects of visual perception." In: *Psychological review* 61.3, page 183.

Barlow, H. (1961). "Possible principles underlying the transformation of sensory messages". In: *Sensory Communication, MIT Press*.

Bordelon, Blake and Cengiz Pehlevan (2021). "Population Codes Enable Learning from Few Examples By Shaping Inductive Bias". In: *bioRxiv*.

Chalk, Matthew, Olivier Marre, and Gasper Tkacik (2018). "Toward a unified theory of efficient, predictive, and sparse coding". In: *Proceedings of the National Academy of Sciences* 115.1, pages 186–191.

Chen, Ting et al. (2020). "A simple framework for contrastive learning of visual representations". In: *arXiv preprint arXiv:2002.05709*.

Chestek, Cynthia A et al. (2007). "Single-neuron stability during repeated reaching in macaque premotor cortex". In: *Journal of Neuroscience* 27.40, pages 10742–10750.

Deitch, Daniel, Alon Rubin, and Yaniv Ziv (Oct. 2021). "Representational drift in the mouse visual cortex". In: *Current Biology* 31.10, pages 1–13.

Driscoll, Laura N et al. (2017). "Dynamic reorganization of neuronal activity patterns in parietal cortex". In: *Cell* 170.5, pages 986–999.

— (2020). "Data from: Dynamic reorganization of neuronal activity patterns in parietal cortex dataset," in: *Dryad, Dataset*. URL: https://doi.org/10.5061/dryad.gqnk98sjq.

Druckmann, Shaul and Dmitri B Chklovskii (2012). "Neuronal circuits underlying persistent representations despite time varying activity". In: *Current Biology* 22.22, pages 2095–2103.

Duffy, Alison et al. (2019). "Variation in sequence dynamics improves maintenance of stereotyped behavior in an example from bird song". In: *Proceedings of the National Academy of Sciences* 116.19, pages 9592–9597.

Földiak, Peter (1990). "Forming sparse representations by local anti-Hebbian learning". In: *Biological cybernetics* 64.2, pages 165–170.

Gallego, Juan A et al. (2020). "Long-term stability of cortical population dynamics underlying consistent behavior". In: *Nature Neuroscience*, pages 1–11.

Gonzalez, Walter G et al. (2019). "Persistence of neuronal representations through time and damage in the hippocampus". In: *Science* 365.6455, pages 821–825.

Harvey, Christopher D, Philip Coen, and David W Tank (2012). "Choice-specific sequences in parietal cortex during a virtual-navigation decision task". In: *Nature* 484.7392, pages 62–68.

Hateren, Johannes H van (1992). "A theory of maximizing sensory information". In: *Biological cybernetics* 68.1, pages 23–29.

Hazan, Liran and Noam E Ziv (2020). "Activity dependent and independent determinants of synaptic size diversity". In: *Journal of Neuroscience* 40.14, pages 2828–2848.

Hubel, David H (1995). *Eye, brain, and vision.* Scientific American Library/Scientific American Books.

Hunter, Gary L et al. (2011). "Tracking rotational diffusion of colloidal clusters". In: *Optics express* 19.18, pages 17189–17202.

Kämmerer, Stefan, Walter Kob, and Rolf Schilling (1997). "Dynamics of the rotational degrees of freedom in a supercooled liquid of diatomic molecules". In: *Physical Review E* 56.5, page 5450.

Kappel, David et al. (2015). "Network plasticity as Bayesian inference". In: *PLoS Comput Biol* 11.11, e1004485.

Katlowitz, Kalman A, Michel A Picardo, and Michael A Long (2018). "Stable sequential activity underlying the maintenance of a precisely executed skilled behavior". In: *Neuron* 98.6, pages 1133–1140.

Kaufman, Matthew T et al. (2014). "Cortical activity in the null space: permitting preparation without movement". In: *Nature neuroscience* 17.3, pages 440–448.

Kinsky, Nathaniel R et al. (2018). "Hippocampal place fields maintain a coherent and flexible map across long timescales". In: *Current Biology* 28.22, pages 3578–3588.

Kriegeskorte, Nikolaus, Marieke Mur, and Peter A Bandettini (2008). "Representational similarity analysis-connecting the branches of systems neuroscience". In: *Frontiers in systems neuroscience* 2, page 4.

Kropff, Emilio and Alessandro Treves (2008). "The emergence of grid cells: Intelligent design or just adaptation?" In: *Hippocampus* 18.12, pages 1256–1269.

Lee, Jae Sung et al. (2020). "The statistical structure of the hippocampal code for space as a function of time, context, and value". In: *Cell* 183.3, pages 620–635.

Li, Ming et al. (2017). "Long-term two-photon imaging in awake macaque monkey". In: *Neuron* 93.5, pages 1049–1057.

Lian, Yanbo and Anthony N Burkitt (2020). "Learning an efficient place cell map from grid cells using non-negative sparse coding". In: *bioRxiv*.

Liberti, William A et al. (2016). "Unstable neurons underlie a stable learned behavior". In: *Nature neuroscience* 19.12, pages 1665–1671.

Luo, Thomas Zhihao et al. (2020). "An approach for long-term, multi-probe Neuropixels recordings in unrestrained rats". In: *Elife* 9, e59716.

Marks, Tyler D. and Michael J. Goard (2021). "Stimulus-dependent representational drift in primary visual cortex". In: *Nature Communications* 12.1, page 5169.

Mau, William, Michael E Hasselmo, and Denise J Cai (2020). "The brain in motion: How ensemble fluidity drives memory-updating and flexibility". In: *Elife* 9, e63550.

Mazza, Marco G et al. (2006). "Relation between rotational and translational dynamic heterogeneities in water". In: *Physical Review Letters* 96.5, page 057803.

Moser, May-Britt, David C Rowland, and Edvard I Moser (2015). "Place cells, grid cells, and memory". In: *Cold Spring Harbor perspectives in biology* 7.2, a021808.

O'Keefe, John and Jonathan Dostrovsky (1971). "The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat." In: *Brain research*.

Olshausen, Bruno A. and David J. Field (1997). "Sparse coding with an overcomplete basis set: A strategy employed by V1?" In: *Vision Research* 37.23, pages 3311–3325. ISSN: 0042-6989. DOI: https://doi.org/10.1016/S0042-6989(97)00169-7.

Pehlevan, Cengiz (2019). "A spiking neural network with local learning rules derived from nonnegative similarity matching". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 7958–7962.

Pehlevan, Cengiz and Dmitri B Chklovskii (2014). "A Hebbian/anti-Hebbian network derived from online non-negative matrix factorization can cluster and discover sparse features". In: *2014 48th Asilomar Conference on Signals, Systems and Computers*. IEEE, pages 769–775.

— (2019). "Neuroscience-inspired online unsupervised learning algorithms: Artificial neural networks". In: *IEEE Signal Processing Magazine* 36.6, pages 88–96.

Pehlevan, Cengiz, Tao Hu, and Dmitri B Chklovskii (2015). "A hebbian/anti-hebbian neural network for linear subspace learning: A derivation from multidimensional scaling of streaming data". In: *Neural computation* 27.7, pages 1461–1495.

Pehlevan, Cengiz, Sreyas Mohan, and Dmitri B Chklovskii (2017). "Blind nonnegative source separation using biological neural networks". In: *Neural computation* 29.11, pages 2925–2954.

Pehlevan, Cengiz, Anirvan M Sengupta, and Dmitri B Chklovskii (2018). "Why do similarity matching objectives lead to hebbian/anti-hebbian networks?" In: *Neural computation* 30.1, pages 84–124.

Peña, José Luis and Masakazu Konishi (2001). "Auditory spatial receptive fields created by multiplication". In: *Science* 292.5515, pages 249–252.

Rajan, Kanaka, Christopher D Harvey, and David W Tank (2016). "Recurrent network models of sequence generation and memory". In: *Neuron* 90.1, pages 128–142.

Rao, Rajesh PN and Dana H Ballard (1999). "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects". In: *Nature neuroscience* 2.1, pages 79–87.

Rokni, Uri et al. (2007). "Motor learning with unstable neural representations". In: *Neuron* 54.4, pages 653–666.

Rubin, Alon et al. (2019). "Revealing neural correlates of behavior without behavioral measurements". In: *Nature communications* 10.1, pages 1–14.

Rule, Michael Everett, Adrianna R Loback, et al. (2020). "Stable task information from an unstable neural population". In: *eLife* 9, e51121.

Rule, Michael Everett and Timothy O'Leary (2021). "Self-Healing Neural Codes". In: *bioRxiv*.

Rule, Michael Everett, Timothy O'Leary, and Christopher D Harvey (2019). "Causes and consequences of representational drift". In: *Current opinion in neurobiology* 58, pages 141–147.

Rumpel, Simon and Jochen Triesch (2016). "The dynamic connectome". In: *Neuroforum* 22.3, pages 48–53.

Samorodnitsky, Gennady and Murad S Taqqu (2017). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance: Stochastic Modeling*. Routledge.

Schoonover, Carl E. et al. (2021). "Representational drift in primary olfactory cortex". In: *Nature* 594.7864, pages 541–546. DOI: 10.1038/s41586-021-03628-7.

Sengupta, Anirvan M et al. (2018). "Manifold-tiling localized receptive fields are optimal in similarity-preserving neural networks". In: *Advances in Neural Information Processing Systems*, pages 7080–7090.

Solstad, Trygve, Edvard I Moser, and Gaute T Einevoll (2006). "From grid cells to place cells: a mathematical model". In: *Hippocampus* 16.12, pages 1026–1031.

Srinivasan, Mandyam Veerambudi, Simon Barry Laughlin, and Andreas Dubs (1982). "Predictive coding: a fresh view of inhibition in the retina". In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 216.1205, pages 427–459.

Stensola, Hanne et al. (2012). "The entorhinal grid map is discretized". In: *Nature* 492.7427, pages 72–78.

Ulivi, Alessandro F et al. (2019). "Longitudinal two-photon imaging of dorsal hippocampal CA1 in live mice". In: *JoVE (Journal of Visualized Experiments)* 148, e59598.

Yoon, KiJung et al. (2016). "Grid cell responses in 1D environments assessed as slices through a 2D lattice". In: *Neuron* 89.5, pages 1086–1099.

Zbontar, Jure et al. (2021). "Barlow twins: Self-supervised learning via redundancy reduction". In: *arXiv preprint arXiv:2103.03230*.

Ziv, Yaniv et al. (2013). "Long-term dynamics of CA1 hippocampal place codes". In: *Nature neuroscience* 16.3, page 264.

Zuo, Yi et al. (2005). "Long-term sensory deprivation prevents dendritic spine loss in primary somatosensory cortex". In: *Nature* 436.7048, pages 261–265.

Supplemental Material for: Coordinated drift of receptive fields during noisy representation learning

Shanshan Qin, Shiva Farashahi, David Lipshutz, Anirvan M. Sengupta, Dmitri B. Chklovskii, and Cengiz Pehlevan

I. DERIVATION OF THE ROTATIONAL DIFFUSION CONSTANT THE IN THE LINEAR HEBBIAN/ANTI-HEBBIAN NETWORK

In this section, we derive an analytical expression for the rotational diffusion constant defined by [1, 2]

$$D_\varphi \equiv \lim_{t\to\infty} \frac{1}{4t}\langle|\vec{\varphi}(t)-\vec{\varphi}(0)|^2\rangle, \tag{1}$$

where brackets mean averaging over different realizations of the noise. Obtaining an exact expression for D_φ is difficult, but we are able to derive an approximation that matches numerical experiments well, as shown in Figure 2 E, F and G of main text.

Our approach relies on two simplifications. First, we define the single-step angular displacement

$$\Delta\vec{\varphi}_i \equiv \vec{\varphi}(i)-\vec{\varphi}(i-1) \tag{2}$$

and note that

$$\langle|\vec{\varphi}(t)-\vec{\varphi}(0)|^2\rangle = \sum_{i=1}^{t}\langle|\Delta\vec{\varphi}_i|^2\rangle + \sum_{i=1}^{t}\sum_{j=1,i\neq j}^{t}\langle\Delta\vec{\varphi}_i\cdot\Delta\vec{\varphi}_j\rangle. \tag{3}$$

We assume that the correlation between angular displacements at different times is negligible. Therefore, we approximate

$$D_\varphi \approx \lim_{t\to\infty}\frac{1}{4t}\sum_{k=1}^{t}\langle|\Delta\vec{\varphi}_i|^2\rangle. \tag{4}$$

Second, we assume that the network weights start at a configuration that is already an optimal solution to the similarity matching objective, projecting the input to its principal subspace, and the drift keeps the weights in the optimal solution space. This is a reasonable approximation because of a linear stability analysis presented in [3, 4]. We now review that argument. We refer an optimal solution to the similarity matching problem in the offline setting without noise as a fixed point and denote it with a $\hat{}$. We note that a general perturbation of feature map $\delta\mathbf{F}$ around a fixed point $\hat{\mathbf{F}}=\hat{\mathbf{M}}^{-1}\hat{\mathbf{W}}$ can be decomposed as

$$\delta\mathbf{F} = \delta\mathbf{A}\hat{\mathbf{F}} + \delta\mathbf{S}\hat{\mathbf{F}} + \delta\mathbf{B}\hat{\mathbf{G}}, \tag{5}$$

where $\delta\mathbf{A}$ is a $k\times k$ antisymmetric matrix, $\delta\mathbf{S}$ is a $k\times k$ symmetric matrix, and $\hat{\mathbf{G}}$ is a $(n-k)\times n$ matrix with orthonormal rows. These rows are chosen to be orthogonal to the rows of $\mathbf{F}$. $\delta\mathbf{B}$ is a $k\times(n-k)$ matrix [4]. So we have $\delta\mathbf{A}+\delta\mathbf{S}=\delta\mathbf{F}\hat{\mathbf{F}}$. The first term corresponds to a rotation of the neural filter basis of the principal subspace, the second term captures deviations from orthogonality of the basis vectors within the subspace, and the third term captures perturbations of the weight vectors that lead to projecting outside the principal subspace. As shown in [4], the fixed point is stable to the perturbation due to the second and third term, meaning they decay exponentially to zero, making a principal subspace projection linearly stable. Therefore, we consider drift due to the first term, which rotates neural filters and, in turn, the data cloud. We find that (see below) in this limit, $\langle|\Delta\vec{\varphi}_i|^2\rangle$ is independent of time step i. Therefore, our final approximation is

$$D_\varphi \approx \frac{1}{4\Delta t}\langle|\Delta\vec{\varphi}|^2\rangle, \tag{6}$$

where Δt is the small time interval elapsed during a single step update, and $\Delta\vec{\varphi}$ arises from a noisy synaptic update to the network with an optimal set of synapses. This quantity is called mean squared angular displacement (MSAD). This approximation turns out to match simulations very well as shown in Figure 2 E, F and G.

Next, we calculate D_φ. In the linear Hebbian/anti-Hebbian network for principle subspace projection task, the learning rule with synaptic noise is

$$\Delta\mathbf{W} = \eta(\mathbf{y}_t\mathbf{x}_t^\top-\mathbf{W})+\boldsymbol{\xi}^W, \quad \Delta\mathbf{M} = \eta(\mathbf{y}_t\mathbf{y}_t^\top-\mathbf{M})+\boldsymbol{\xi}^M, \tag{7}$$

where $\langle \xi_{ij}^W(t)\rangle = \langle \xi_{ij}^M(t)\rangle = 0$ and $\langle (\xi_{ij}^W(t)\xi_{kl}^W(t')\rangle = \eta\sigma_1^2\delta_{ik}\delta_{jl}\delta(t-t')$, $\langle \xi_{ij}^M(t)\xi_{kl}^M(t')\rangle = \eta\sigma_2^2\delta_{ik}\delta_{jl}\delta(t-t')$. As discussed, by estimating the variance of the rotation the learned representation during a single-step update under rule (7), we can define an effective rotational diffusion constant that is related to this variance. More specifically, in the small update and noise regime, $\delta\mathbf{A}$ is related to an infinitesimal rotation $\mathbf{R}$ by $\mathbf{R} = \exp\delta\mathbf{A} = \exp(\vec{\theta}\cdot\vec{\mathbf{L}})$, where $\vec{\mathbf{L}}$ is the infinitesimal rotation generator [5]. $\vec{\mathbf{L}}$ is a tensor, whose components can be written in matrix form.

We start by writing $\delta\mathbf{F}$ in terms of the perturbation of $\hat{\mathbf{W}}, \hat{\mathbf{M}}$:

$$\delta\mathbf{F} = \hat{\mathbf{M}}^{-1}\delta\mathbf{W} - \hat{\mathbf{M}}^{-1}\delta\mathbf{M}\hat{\mathbf{M}}^{-1}\hat{\mathbf{W}} = \hat{\mathbf{M}}^{-1}(\delta\mathbf{W} - \delta\mathbf{M}\hat{\mathbf{F}}), \tag{8}$$

where we have used the property $\hat{\mathbf{F}}\hat{\mathbf{F}}^\top = \mathbf{I}$. Right-multiplying (8) by $\hat{\mathbf{F}}$ and using (7), we have

$$\delta\mathbf{F}\hat{\mathbf{F}}^\top = \hat{\mathbf{M}}^{-1}(\delta\mathbf{W}\hat{\mathbf{F}}^\top - \delta\mathbf{M}) = \hat{\mathbf{M}}^{-1}\left(\eta(\mathbf{y}\mathbf{x}_t^\top - \hat{\mathbf{W}})\hat{\mathbf{F}}^\top - \eta(\mathbf{y}\mathbf{y}^\top - \hat{\mathbf{M}}) + \boldsymbol{\xi}^W\hat{\mathbf{F}}^\top - \boldsymbol{\xi}^M\right) = \hat{\mathbf{M}}^{-1}(\boldsymbol{\xi}^W\hat{\mathbf{F}}^\top - \boldsymbol{\xi}^M), \tag{9}$$

where we have used the fact

$$\hat{\mathbf{M}}^{-1}\left(\mathbf{y}_t\mathbf{x}_t^\top - \hat{\mathbf{W}})\hat{\mathbf{F}}^\top - (\mathbf{y}_t\mathbf{y}_t^\top - \hat{\mathbf{M}})\right) = \hat{\mathbf{M}}^{-1}\left(\mathbf{y}_t\mathbf{y}_t^\top - \hat{\mathbf{W}}\hat{\mathbf{F}}^\top - \mathbf{y}_t\mathbf{y}_t^\top + \hat{\mathbf{M}}\right) = -\hat{\mathbf{M}}^{-1}\hat{\mathbf{W}}\hat{\mathbf{F}}^\top + \mathbf{I} = \mathbf{0} \tag{10}$$

Now, the antisymmetric part $\delta\mathbf{A} = \frac{1}{2}(\delta\mathbf{F}\hat{\mathbf{F}}^\top - \hat{\mathbf{F}}\delta\mathbf{F}^\top)$ can be written down explicitly:

$$\delta\mathbf{A} = \frac{1}{2}[(\hat{\mathbf{M}}^{-1}\boldsymbol{\xi}^W\hat{\mathbf{F}}^\top - \hat{\mathbf{F}}\boldsymbol{\xi}^{W\top}\hat{\mathbf{M}}^{-1}) + (\boldsymbol{\xi}^{M\top}\hat{\mathbf{M}}^{-1} - \hat{\mathbf{M}}^{-1}\boldsymbol{\xi}^M)]. \tag{11}$$

The mean squared angular displacement (MSAD) is related to $\delta\mathbf{A}$. To see this more clearly, consider a d-dimension rotation, which can be interpreted as rotation in a $d-1$ dimensional hyperplane from one unit vector to another unit vector. Given any two d-dimensional orthogonal unit vector $\mathbf{e_1}, \mathbf{e_2}$, i.e., $\mathbf{e_1}^\top\cdot\mathbf{e}_1 = \mathbf{e_2}^\top\cdot\mathbf{e}_2 = 1$, $\mathbf{e_1}^\top\cdot\mathbf{e}_2 = \mathbf{e_2}^\top\cdot\mathbf{e}_1 = 0$. The generator for this rotation can be represented as

$$L_{\mathbf{e_1e_2}} = \mathbf{e}_2\mathbf{e}_1^\top - \mathbf{e}_1\mathbf{e}_2^\top. \tag{12}$$

Hence $\delta\mathbf{A}$ can be expressed as $\delta\mathbf{A} = \Delta\varphi L_{\mathbf{e_1e_2}}$, with $\Delta\varphi$ reflecting the rotation 'amplitude'. Using the fact that $\mathrm{Tr}(L_{\mathbf{e_1e_2}}L_{\mathbf{e_1e_2}}^\top) = 2$, we have

$$2(\Delta\varphi)^2 = \mathrm{Tr}(\delta\mathbf{A}\delta\mathbf{A}^\top). \tag{13}$$

The variance of δA_{ij} is

$$\langle\delta A_{ij}^2\rangle = \frac{\eta}{4}\left[\sigma_1^2\sum_{kl}(\tilde{M}_{ik}F_{jl} - \tilde{M}_{jk}F_{il})^2 + \sigma_2^2\sum_k(\tilde{M}_{kj}^2 + \tilde{M}_{ki}^2) - 2\delta_{ij}\tilde{M}_{ki}\tilde{M}_{kj}\right]\Delta t, \tag{14}$$

where $\tilde{\mathbf{M}} \equiv \hat{\mathbf{M}}^{-1}$ and the average $\langle\rangle$ is over the noise distribution, Δt is the time interval of the single-step update.

Using the fact that $\mathrm{eig}(\hat{\mathbf{M}}) = [\lambda_1, \cdots, \lambda_k]$ [4], and $\mathrm{Tr}\tilde{M} = \sum_{i=1}^k 1/\lambda_i^2$. We have

$$\langle\mathrm{Tr}\delta\mathbf{A}\delta\mathbf{A}^\top\rangle = \sum_{ij}\langle\delta A_{ij}^2\rangle = \frac{1}{2}\eta(k-1)\Delta t(\sigma_1^2+\sigma_2^2)\sum_{i=1}^k\frac{1}{\lambda_i^2}, \tag{15}$$

where we have used the fact. We then define the rotational diffusion constant D_φ by the relation $\langle|\varphi(t+\Delta t)-\varphi(t)|^2\rangle = 2(k-1)D_\varphi\Delta t$. With Eq.(13) and Eq.(15), we arrive Eq. 5 in the main text.

II. DERIVATION OF THE EFFECTIVE DIFFUSION CONSTANT IN THE RING MODEL

Here, we calculate the diffusion constant in the ring model for a single output neuron, using the MSAD approximation as before.

We start from the simplest setup for the 1D place cell model: a single place cell which receives input from the "ring" manifold, i.e., the position is parameterized as $\mathbf{x} = [\cos\theta, \sin\theta]^\top$. The response of the neurons is given by

$$y(\theta) = \frac{1}{m+\beta}[w_1\cos\theta + w_2\sin\theta - \alpha b]_+. \tag{16}$$

Here and after, we use $[x]_+$ to denote the rectified linear function. We define a steady state where the average update to the weights is zero. Denoting the stationary state weights as $\{w_1^*, w_2^*, m^*, b^*\}$, this leads to the conditions:

$$\langle w_1^* \rangle = \langle y(\theta)\cos\theta \rangle_\theta, \quad \langle w_2^* \rangle = \langle y(\theta)\sin\theta \rangle_\theta, \quad \langle m^* \rangle = \langle y^2(\theta) \rangle_\theta, \quad \langle b^* \rangle = \alpha \langle y(\theta) \rangle_\theta, \tag{17}$$

Where $\langle \cdot \rangle_\theta$ means averaging over $\theta \in [-\pi, \pi)$ which is a uniform distribution. These equations can be solved self-consistently using an ansatz of the form:

$$y_\phi(\theta) = \mu[\cos(\theta - \phi) - \cos(\psi)]_+, \quad \theta \in [-\pi, \pi] \tag{18}$$

where ψ determines the "width" of the RF, and $\mu(1 - \cos\psi)$ is the peak amplitude and ϕ is the centroid of the receptive field. Plugging (18) into (16) and (17), we find that

$$\langle w_1^* \rangle = \frac{\mu}{4\pi}(2\psi - \sin 2\psi)\cos\phi, \tag{19}$$

$$\langle w_2^* \rangle = \frac{\mu}{4\pi}(2\psi - \sin 2\psi)\sin\phi, \tag{20}$$

$$\langle m^* \rangle = \frac{\mu^2}{4\pi}(4\psi + 2\psi\cos 2\psi - 3\sin 2\psi), \tag{21}$$

$$\langle b^* \rangle = \frac{\alpha\mu}{\pi}(\sin\psi - \psi\cos\psi). \tag{22}$$

(16) can be rewritten as

$$y(\theta) = \frac{\sqrt{w_1^2 + w_2^2}}{m + \beta}\left[\frac{w_1}{\sqrt{w_1^2 + w_2^2}}\cos\theta + \frac{w_2}{\sqrt{w_1^2 + w_2^2}}\sin\theta - \frac{\alpha b}{\sqrt{w_1^2 + w_2^2}}\right]_+ \tag{23}$$

Compared with (18), we have

$$\mu = \frac{\sqrt{w_1^{*2} + w_2^{*2}}}{m^* + \beta}, \quad \alpha b^* = \sqrt{w_1^{*2} + w_2^{*2}}\cos\psi. \tag{24}$$

Combining (19)-(22) and (24), we get the dependence of μ and ψ on α, β, given parametrically by

$$\mu^2 = \frac{2\psi - \sin 2\psi - 4\beta\pi}{4\psi + 2\psi\cos 2\psi - 3\sin 2\psi}, \quad \alpha^2 = \frac{\cos\psi(2\psi - \sin 2\psi)}{4(\sin\psi - \psi\cos\psi)}. \tag{25}$$

Next, we proceed to estimate the drift due to noisy synaptic updates. From (23), we have

$$w_1^*\cos\phi + w_2^*\sin\phi = \sqrt{w_1^{*2} + w_2^{*2}} = \frac{\mu}{4\pi}(2\psi - \sin 2\psi) \equiv \hat{\mu}, \tag{26}$$

where we have defined $\hat{\mu}$ to simplify the following notations. Using the fact that $dy(\theta)/d\theta = 0$ at $\theta = \phi$, we have $\tan(\phi) = w_2^*/w_1^*$ and

$$\frac{d\phi}{dt} = \frac{1}{\hat{\mu}}\left(\frac{dw_2}{dt}\cos\phi - \frac{dw_1}{dt}\sin\phi\right). \tag{27}$$

We are interested in how the centroid of the RF changes when a perturbation is added to the stationary weight vector

$$\Delta\phi = \frac{1}{\hat{\mu}}(\Delta w_2\cos\phi - \Delta w_1\sin\phi), \tag{28}$$

$$\Delta w_1 = \eta(y(\theta)\cos\theta - w_1^*) + \xi_1, \tag{29}$$

$$\Delta w_2 = \eta(y(\theta)\sin\theta - w_2^*) + \xi_2. \tag{30}$$

The Gaussian white noise terms have the following property: $\langle \xi_1 \rangle = \langle \xi_2 \rangle = 0, \langle \xi_1^2 \rangle = \langle \xi_2^2 \rangle = \eta\sigma^2\Delta t$ with Δt as the time interval between two adjacent update events. From (26), we have $w_1^* = \hat{\mu}\cos\phi, w_2^* = \hat{\mu}\sin\phi$. Then $\Delta\phi$ can be written as

$$\begin{aligned}\Delta\phi &= \frac{1}{\hat{\mu}}\left(\eta\mu[\cos(\theta - \phi) - \cos\psi]_+\sin\theta\cos\phi - \eta\mu[\cos(\theta - \phi) - \cos\psi]_+\cos\theta\sin\phi + (w_1^*\xi_2 - w_2^*\xi_1)\right) \\ &= \frac{1}{\hat{\mu}}\{\eta\mu[\cos(\theta - \phi) - \cos\psi]_+\sin(\theta - \phi) + (\xi_2\cos\phi - \xi_1\sin\phi)\} \\ &= \frac{1}{\hat{\mu}}\{h_t(\theta, \phi, \psi) - (\xi_2\cos\phi - \xi_1\sin\phi)\},\end{aligned} \tag{31}$$

where we have defined $h_t(\theta, \phi, \psi) = \eta\mu[\cos(\theta - \phi) - \cos\psi]_+$. Since in the online learning, θ is sampled randomly, we can regard h_t as a stochastic process. Averaging over θ, we have

$$\langle(\Delta\phi)^2\rangle = \frac{\mu^2}{\hat{\mu}^2}\eta^2\langle([\cos(\theta - \phi) - \cos\psi]_+^2 \sin^2(\theta - \phi))\rangle_\theta + \frac{1}{\hat{\mu}^2}(\cos^2\phi\langle\xi_2^2\rangle + \sin^2\phi\langle\xi_1^2\rangle) = \gamma\eta^2\Delta t + \frac{\eta\sigma^2}{\hat{\mu}^2}\Delta t, \tag{32}$$

where

$$\gamma \equiv \frac{\mu^2}{\hat{\mu}^2}\frac{1}{2\pi}\int_{-\pi}^{\pi}[\cos(\theta - \phi) - \cos\psi]_+^2 \sin^2(\theta - \phi)d\theta = \frac{\pi}{6}\frac{36\psi + 24\psi\cos(2\psi) - 28\sin(2\psi) - \sin(4\psi)}{(2\psi - \sin(2\psi))^2} \tag{33}$$

Using the relation $\langle(\Delta\phi)^2(\Delta t)\rangle = 2D\Delta t$, we have

$$D \approx \gamma\eta^2 + \frac{\eta\sigma^2}{\hat{\mu}^2}, \tag{34}$$

where we use $\approx$ because we calculate single-step MSAD.

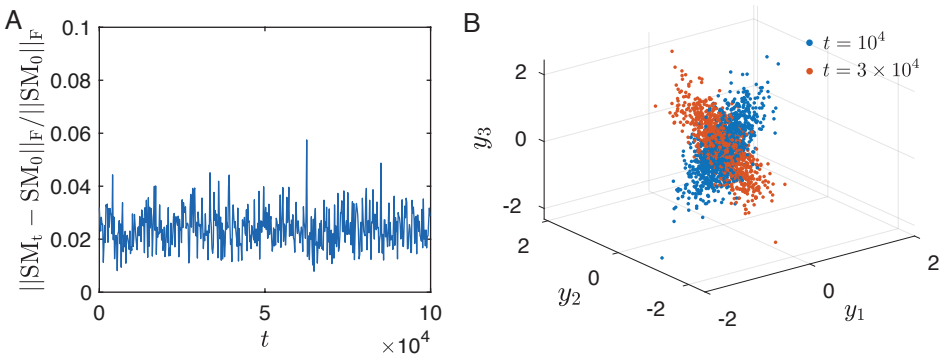

FIG. S1. (A) The relative change of Frobenus norm of the similarity matrix at time t compared with time point 0 in the PSP task. (B) Ensemble of output $\mathbf{Y} \equiv [\mathbf{y}_1, \cdots, \mathbf{y}_1]$ at two time points. The data clouds have ellipsoid shape. Related to figure 2 in the main text. Parameters are the same as in figure 2 of the main text.

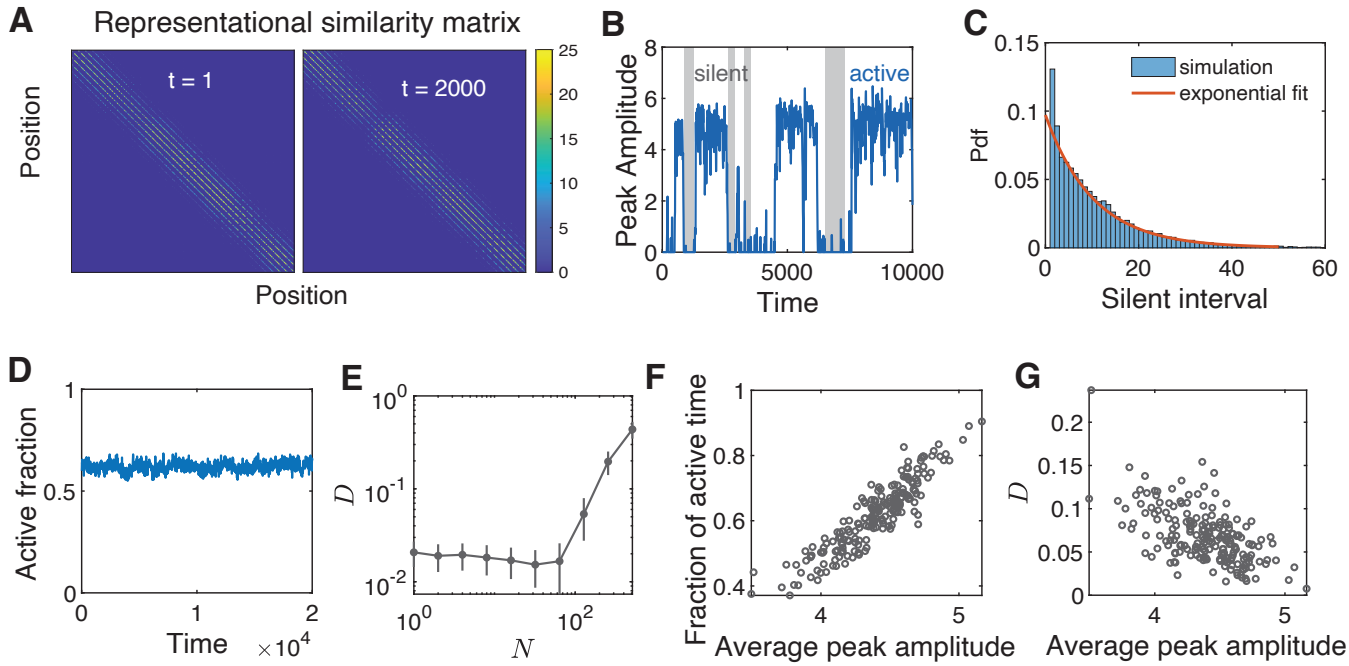

FIG. S2. Drift of 2D place cells in the model. (A) Representational similarity is preserved despite the continuous drift of place cell RFs. Positions on the plane are represented by an index from 1 to 1024. (B) The RFs are intermittent. The peak amplitude of an example place field has active and silent bouts. (C) The interval of silent bouts follow exponential distribution. (D) At population level, there is a constant fraction of active RFs over time. (E) Dependence of effective diffusion constant on the total number output neurons. (F,G) Place cells that have stronger place fields tend to be active more often (F) and also more stable as quantified by smaller diffusion constant (G). Parameters used are the same as Figure 5 in the main text.

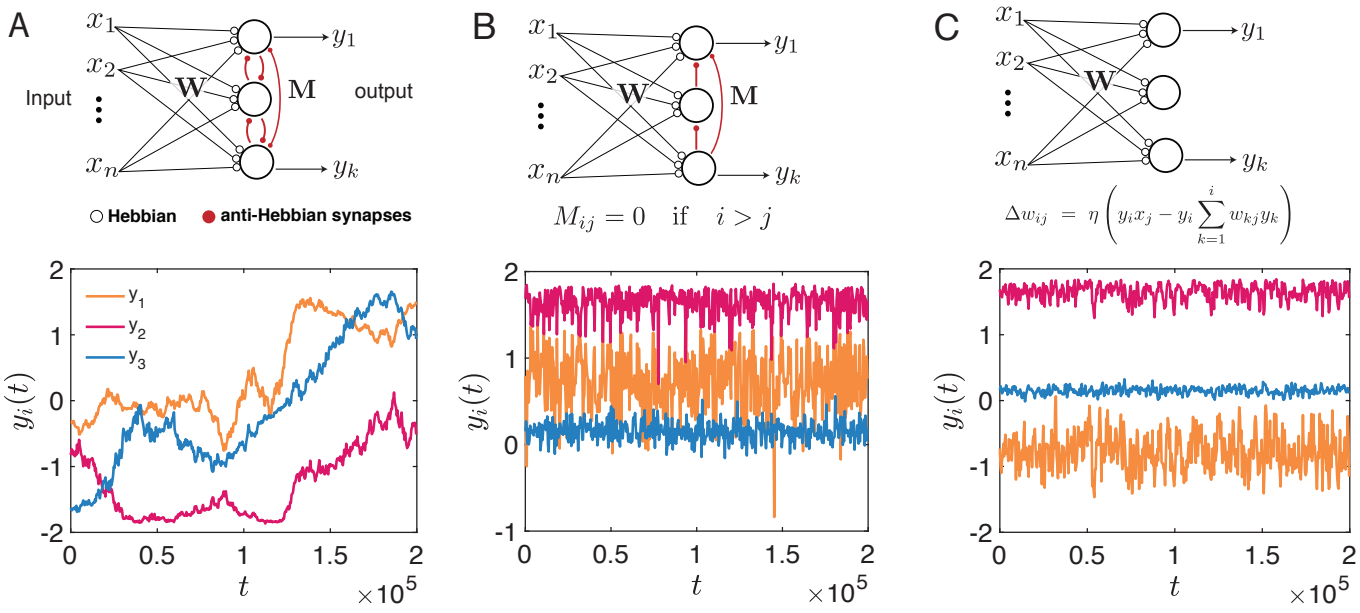

FIG. S3. Degeneracy of learning objective function and representational drift. We compare the long-term behavior of learned representations in three different networks. (A) Upper: the Hebbian/anti-Hebbian network for PSP. Lower: the evolution of the three components of a representation $\mathbf{y}_t$. (B) Upper: The network differs from Hebbian/anti-hebbain network only in the lateral matrix $\mathbf{M}$ which break the rotational symmetry of PSP solution. The learning rule is the same. Lower: the learned representation only fluctuates around a equilibrium. (C) A single feedforward network that perform online principle component analysis with Sanger's rule [6]. This network has only feedforward input matrix $\mathbf{W}$ and the learning rule is nonlocal. Lower: learned representation is relatively stable in the presence of noise. Parameters are the same as in the figure 2 of main text except that $\eta = 0.01$.

[1] S. Kämmerer, W. Kob, and R. Schilling, Dynamics of the rotational degrees of freedom in a supercooled liquid of diatomic molecules, Physical Review E **56**, 5450 (1997).
[2] M. G. Mazza, N. Giovambattista, F. W. Starr, and H. E. Stanley, Relation between rotational and translational dynamic heterogeneities in water, Physical Review Letters **96**, 057803 (2006).
[3] C. Pehlevan and D. Chklovskii, A normative theory of adaptive dimensionality reduction in neural networks, in *Advances in neural information processing systems* (2015) pp. 2269–2277.
[4] C. Pehlevan, A. M. Sengupta, and D. B. Chklovskii, Why do similarity matching objectives lead to hebbian/anti-hebbian networks?, Neural computation **30**, 84 (2018).
[5] A. Zee, *Group theory in a nutshell for physicists*, Vol. 17 (Princeton University Press, 2016).
[6] T. D. Sanger, Optimal unsupervised learning in a single-layer linear feedforward neural network, Neural networks **2**, 459 (1989).