Neuroscience

Representational drift as a result of implicit regularization

Aviv Ratzon ✉, **Dori Derdikman, Omri Barak**

Rappaport Faculty of Medicine, Technion - Israel Institute of Technology, Haifa 31096, Israel • Network Biology Research Laboratory, Technion - Israel Institute of Technology, Haifa 32000, Israel

https://en.wikipedia.org/wiki/Open_access

Copyright information

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
April 11, 2024 (this version)

Reviewed preprint version 1
October 10, 2023

Posted to preprint server
July 2, 2023

Sent for peer review
June 20, 2023

Abstract

Recent studies show that, even in constant environments, the tuning of single neurons changes over time in a variety of brain regions. This representational drift has been suggested to be a consequence of continuous learning under noise, but its properties are still not fully understood. To investigate the underlying mechanism, we trained an artificial network on a simplified navigational task. The network quickly reached a state of high performance, and many units exhibited spatial tuning. We then continued training the network and noticed that the activity became sparser with time. Initial learning was orders of magnitude faster than ensuing sparsification. This sparsification is consistent with recent results in machine learning, in which networks slowly move within their solution space until they reach a flat area of the loss function. We analyzed four datasets from different labs, all demonstrating that CA1 neurons become sparser and more spatially informative with exposure to the same environment. We conclude that learning is divided into three overlapping phases: (i) Fast familiarity with the environment; (ii) slow implicit regularization; (iii) a steady state of null drift. The variability in drift dynamics opens the possibility of inferring learning algorithms from observations of drift statistics.

eLife assessment

This study presents a new and **important** theoretical account of spatial representational drift in the hippocampus. The evidence supporting the claims is **convincing**, with a clear and accessible explanation of the phenomenon. Overall, this study will likely attract researchers exploring learning and representation in both biological and artificial neural networks.

What do we mean when we say that the brain represents the external world? One interpretation is the existence of neurons whose activity is tuned to world variables. Such neurons have been observed in many contexts: place cells [1, 2] – which are tuned to position in a specific context, visual cells [3] – which are tuned to specific visual cues, neurons that are tuned to the execution of actions [4] and more. This tight link between the external world and neural activity might suggest that, in the absence of environmental or behavioral changes, neural activity is constant. In contrast, recent studies show that, even in constant environments, the tuning of single neurons to outside world variables gradually changes over time in a variety of brain regions, even long after good representations of the stimuli were achieved. This phenomenon has

been termed *representational drift*, and has changed the way we think about the stability of memory and perception, but its driving forces and properties are still unknown [5, 6, 7, 8, 9] (see [10, 11] for an alternative account).

There are at least two immediate theoretical questions arising from the observation of drift – why does it happen, and whether and how behavior is resistant to it [12, 13]? One mechanistic explanation is that the underlying anatomical substrates are themselves undergoing constant change, such that drift is a direct manifestation of this structural morphing [14]. A normative interpretation posits that drift is a solution to a computational demand, such as temporal encoding [15], 'drop-out' regularization [16], exploration of the solution space [17], or re-encoding during continual learning [12]. Several studies also address the resistance question, providing possible explanations on how behavior can be robust to such phenomena [18, 19, 20, 21].

Here, we focus on the mechanistic question, and leverage analyses of drift statistics for this purpose. Specifically, recent studies suggest that representational drift in the CA1 is driven by active experience [22, 23]. Namely, rate maps decorrelate more when mice are active for a longer time in a given context. This implies that drift is not just a passive process, but rather an active learning one. As drift seems to occur after an adequate representation has formed, it seems fitting to model it as a form of a continuous learning process.

This approach has been recently explored by [24, 25]. They considered continuous learning in noisy, overparameterized neural networks. Because the system is overparameterized, a manifold of zero-loss solutions exists. [24] showed that for feedforward neural networks (FNNs) trained using Hebbian learning with added parameter noise, units change their tuning over time. This was due to an *undirected* random walk within the manifold of solutions. The coordinated drift of neighboring place fields was used as evidence to support this view. The phenomenon of undirected motion within the space of solutions seems plausible, as all members of this space achieve equally good performance (**Fig 1A** left). However, there may be other properties of the solutions (**Fig 1B**) that vary along this manifold, which could potentially bias drift in a certain direction (**Fig 1A** right). It is likely that the drift observed in experiments is a combination of both an undirected and directed movement. We will now introduce theoretical results from machine learning that support the possibility of directed drift.

Recent work provided a tractable analytical framework for the learning dynamics of Stochastic Gradient Descent (SGD) with added noise and an overparameterized regime [26, 27, 28]. These studies showed that, after the network has converged to the zero-loss manifold, a second-order effect biases the random walk along a specific direction within this manifold. This direction reduces an implicit regularizer, determined by the type of noise the network is exposed to. The regularizer is related to the Hessian of the loss – a measure of the flatness of the loss landscape in the vicinity of the solutions. Since this directed movement is a second-order effect, its timescale is orders of magnitude larger than that of the initial convergence.

Consider a biological neural network performing a task. The machine learning (ML) implicit regularization mentioned above requires three components: an overparameterized regime, noise, and SGD. Both biological and artificial networks possess a large number of synapses, or parameters, and hence can reasonably be expected to be overparameterized. Noise can emerge from the external environment or from internal biological elements. It is not reasonable to assume that a precise form of gradient descent is implemented in the brain [29], thereby casting doubt on the third element. Nevertheless, biologically plausible rules could be considered as noisy versions of gradient descent, as long as there is a coherent improvement in performance [30, 31]. Motivated by this analogy, we explore representational drift in models and experimental data.

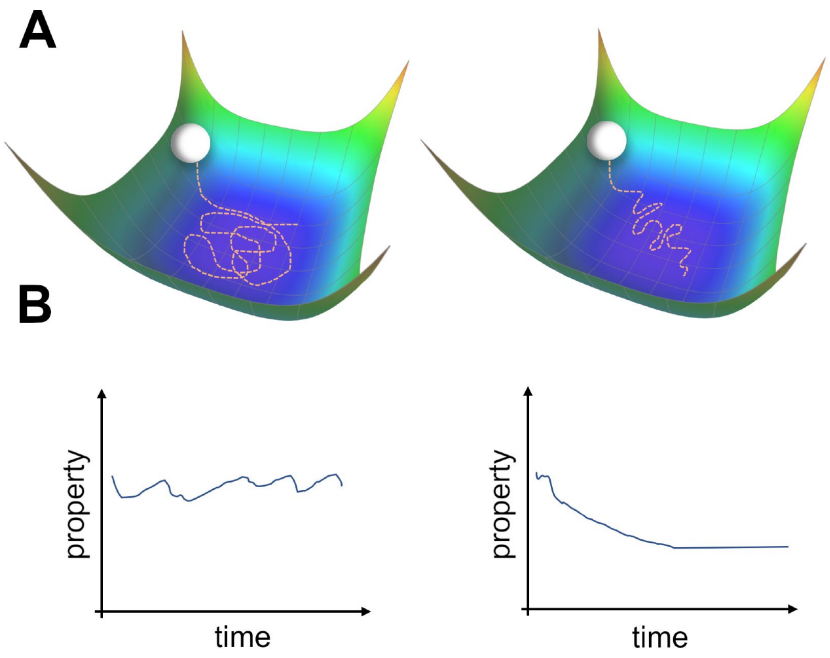


Figure 1

Two types of possible movements within the solution space.

(A) Two options of how drift may look in the solution space. Random walk within the space of equally good solutions that is either undirected (left) or directed (right). (B) The qualitative consequence of the two movement types. For an undirected random walk, all properties of the solution will remain roughly constant (left). For the directed movement there should be a given property that is gradually increasing or decreasing (right).

Because drift is commonly observed in spatially-selective cells, we base our analysis on a model which has been shown to contain such cells [32]. Specifically, we trained artificial neural networks on a predictive coding task in the presence of noise. In this task, an agent moves along a linear track while receiving visual input from the walls, such that the goal is to predict the subsequent input. We observed that hidden layer units became tuned to the latent variable, which is position, in accordance with previous results [32]. We continued training and found that in addition to the gradual change of tuning curves, similar to [24], we witnessed that the fraction of active units decreased slowly while their tuning specificity increased. We show that these results align with experimental observations from the CA1 area - namely that when exposed to a novel environment, the number of active cells reduces while their tuning specificity increases long after the environment is already familiar. Finally, we demonstrated the connection between this sparsificiation effect and changes to the Hessian, in accordance with ML theory. That is, changes in activity statistics (sparseness) and representations (drift) are the signatures of the movement in solution space to a flatter area, until flatness saturates. We conclude that learning is divided into three overlapping phases: (i) Fast familiarity with the environment in which representations form; (ii) slow implicit regularization which can be recognized by changes in activity statistics; (iii) a steady state of null drift in which representations gradually change.

Results

Spontaneous sparsification in a predictive coding network

To model representational drift in the CA1 area, we chose a simple model that could give rise to spatially- tuned cells [32]. In this model, an agent traverses a corridor while slightly modulating its angle with respect to the main axis (**Fig 2A**). The walls are textured by a fixed smooth noisy signal, and the agent receives this as input according to its current field of view. The model itself is a single hidden layer feedforward network, with the velocity and visual field as inputs. The desired output is the predicted visual input in the next time step. The model equations are given by:

$$\hat{\mathbf{y}}_t = \sigma(\mathbf{x}_t\mathbf{m}^T + \mathbf{b})\mathbf{n}^T, \quad (1)$$

$$\sigma(x) = max(0, x), \quad (2)$$

where **m** and **n** are the input and output matrices respectively, **b** is the bias vector, and σ is the ReLU activation function. The vector $\sigma(\mathbf{x}_t\mathbf{m}^T + \mathbf{b})$ constitutes the hidden layer, and each element in it is the activation of a given unit for the input at time t. The task is for the network's output, **y**, to match the visual input, **x** of the following time step, resulting in the following loss function:

$$f(\mathbf{m}, \mathbf{n}, \mathbf{b}) = \mathbb{E}_t(\hat{\mathbf{y}}_t - \mathbf{x}_{t+1})^2. \quad (3)$$

We train the network using Gradient Descent (GD), while adding update noise to the learning dynamics:

$$\theta_{\tau+1} = \theta_\tau - \eta\frac{\partial f(\theta_\tau)}{\partial\theta_\tau} + \xi_\tau^{update}, \quad (4)$$

where θ = **(m, n, b)** is the vectorized parameters-vector, τ is the current training step and ξ_τ^{update} is Gaussian noise. We let the network converge to a good solution, demonstrated by a loss plateau, and continue training for an additional period. Note that this additional period can be orders of magnitude longer than the initial training period. The network quickly converged to a low loss and stayed at the same loss during the additional training period (**Fig 2B**). Surprisingly, when

looking at the activity of units within the hidden layer, we noticed that it slowly became sparse (see methods for definitions of sparseness). This sparsification did not hurt performance, because individual units became more informative (**Fig 2C**), as quantified by the average Spatial Information (SI, see methods). When looking at the rate maps of units, i.e. their tuning to position, one can observe an image similar to representational drift observed in experiments [5] – namely that neurons changed their tuning over time (**Fig 2D**). Additionally, their tuning specificity increased in accordance with the SI increase. By observing the correlation matrix of the rate maps over time, it is apparent that there was a gradual change that slowed down (**Fig 2E**). To summarize, we observed a spontaneous sparsification over a timescale much longer than the initial convergence, without introducing any explicit regularization.

At first glance, these results might seem inconsistent with the experimental descriptions of drift reported in the literature in which all metrics are stationary while only representations change [5]. We suggest that there exists an intermediate phase between initial familiarity and stationary activity metrics, which is consistent with the notion of drift, that exhibits a gradual change in activity statistics. It seems that most experimental paradigms require a long pre-exposure, longer than needed to become fully familiarized with the environment, thus missing the suggested effect. We thus analyzed four datasets, from four different labs, in which we believe that the familiarization stage was shorter than in other studies. Some of the analyses were present in the original papers, and others are novel using publicly available data. Three experiments start from a novel environment and one from a relatively short pre-exposure. As shown in **Fig 3**, all datasets are consistent with our simulations - namely that the fraction of active cells reduces while the mean SI per cell increases over a long timescale. See Methods section for a full description of the data sets and analyses, along with another paper in which activity statistics are stationary, for comparison.

Generality of the phenomenon

The theoretical considerations [26, 27], simulation results and experimental results from multiple labs all suggest very general and robust phenomenona.

To explore the sensitivity of our results to specific modeling choices, we systematically varied many of them. Specifically, we replaced tasks (see below) and simulated different activation functions. Perhaps most important, we varied the learning rules, as SGD is not a biologically plausible one. We used both Adam [36] and RMSprop [37], from the ML literature. We also used Stochastic Error-Descent (SED) [38], which does not require gradient calculation and is more biologically plausible (6). For SGD, we also ran simulations with label noise instead of update noise. Finally, we replaced the task with either a simplified predictive coding, random mappings or smoothed random mapping. The motivation for different tasks is twofold. First, there is some arbitrariness in the predictive task we chose. Second, the interpretation of drift as a result of continuous learning in the presence of noise, suggests that this effect goes beyond the specific phenomenon of place cells. That is, drift within the solution space should occur in every type of task and scenario, and could be identified outside the scope of spatial representations. All 1151 simulations, except a negligible few, demonstrated an initial, fast, phase of convergence to low loss, followed by a much slower phase of directed random motion within the low-loss space.

The results of the simulations supported our main conclusion – sparsification dynamics were not sensitive to most of the parameters. For almost all of the simulations, excluding SGD with label noise, the fraction of active units gradually reduced, long after the loss converged (**Fig 4A**, **Fig S1** for further breakdown). For label noise, a slow directed effect was observed but the dynamics were qualitatively different – as predicted by theory [27] and explained in the next section. The fraction of active units did not reduce as much, but the activity of the units did sparsify (**Fig S2**). One qualitative difference observed between simulations was that the timescales could vary by orders of magnitude as a function of the noise scale (**Fig 4B** bottom, see

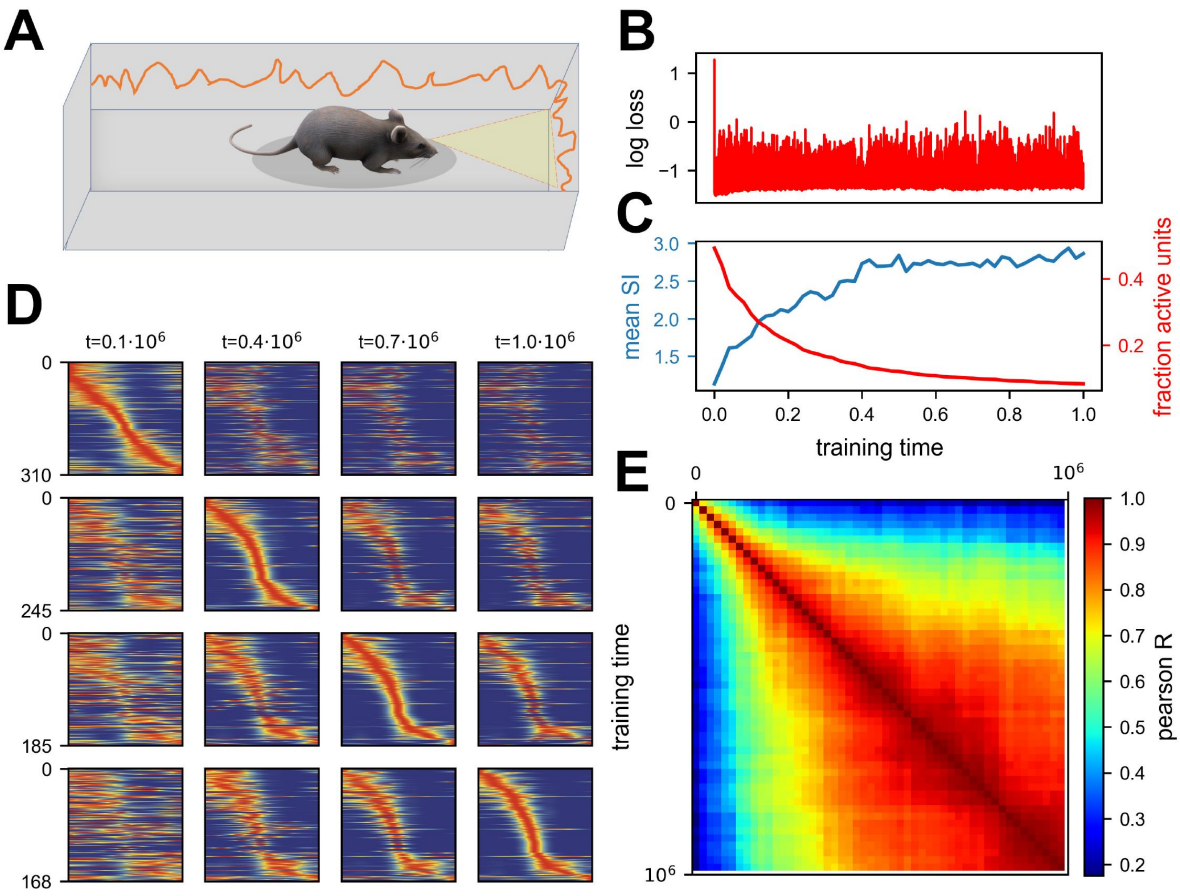


Figure 2

Continuous noisy learning leads to drift and spontaneous sparsification.

(A) Illustration of an agent in a corridor receiving high-dimensional visual input from the walls. (B) Loss as a function of training steps (log scale). Zero loss corresponds to a mean estimator. Note the rapid drop in loss at the beginning, after which it remains roughly constant. (C) Mean spatial information (SI, blue) and fraction of units with non-zero activation for at least one input (red) as a function of training steps. (D) Rate maps sampled at four different time points (columns). Maps in each row are sorted according to a different time point. Sorting is done based on the peak tuning value to the latent variable. (E) Correlation of rate maps between different time points along training. Only active units are used.

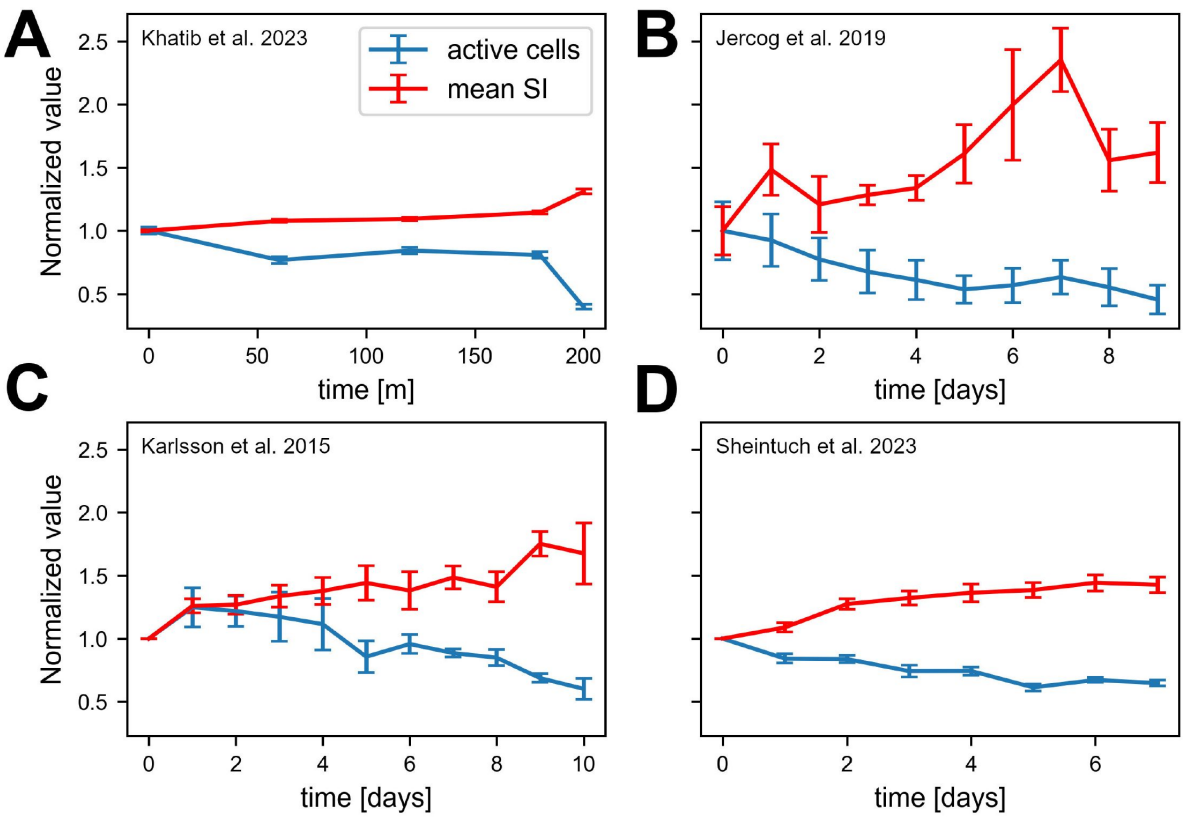


Figure 3

Experimental data consistent with simulations.

Data from four different labs show sparsification of CA1 spatial code, along with an increase in the information of active cells. Values are normalized to the first recording session in each experiment. Error bars show standard error of the mean. (A) Fraction of place cells and mean SI per animal over 200 minutes [22]. (B) Number of cells per animal and mean SI over all cells pooled together over 10 days. Note that we calculated the number of active cells rather than fraction of place cells because of the nature of the available data [33]. (C) Fraction of place cells and mean SI per animal over 11 days [34]. (D) Fraction of place cells and mean SI per animal over 8 days [35].

methods for details). Additionally, apart from simulations that did not converge due to too large timescales, the final sparsity was the same for all networks with the same parameters (**Fig 4B** top), in accordance with results from [24]. In a sense, once noise is introduced, the network is driven to maximal sparsification in a stochastic manner. For Adam, RMSprop and SED sparsification ensued in the absence of any added noise. For SED the explanation is straightforward, as the parameter updates are driven by noise. For Adam and RMSprop, we suggest that in the vicinity of the zero-loss manifold, the second moment acts as noise. In some cases, the networks quickly collapsed to a sparse solution, most likely as a result of the learning rate being too high, in relation to the input statistics [39]. Importantly, for GD without noise, there was no change after the initial convergence.

As a further test of the generality of this phenomenon, we consider the recent simulation from [24] in which representational drift was shown. The learning rule used in this work was very different from the ones we applied, and more biologically plausible. We simulated that network using the published code and found the same type of dynamics as shown above. Namely, the network initially converged to a good solution followed by a longer period of sparsification (**Fig 4C**). Note that in the original publication [24] the focus was on the stage following this sparsification, in which the network indeed maintained a constant fraction of active cells.

In conclusion, we see that noisy learning leads to three phases under rather general conditions. The first phase is the learning of the task and convergence to the manifold of low-loss solutions. The second phase is directed movement on this manifold, driven by a second-order effect of implicit regularization. The third phase is an undirected random walk within the sub-manifold of low loss and maximum regularization. **Table 1** summarizes these three phases and their signature in network metrics. It is important to note that these are not consecutive phases but rather overlapping ones – they occur simultaneously, but due to their different time scales one can identify when each phase is dominant. A rough delineation of when each phase is dominant can be seen in **Fig 4C** background colors – in the first phase (red) the loss function converged, in the second phase (yellow) the fraction of active units reduced substantially. In the third phase (red) both were stationary but the tuning of units continued to change, as shown in the original paper [24]. We speculate that most experimental studies about drift demonstrated only the third phase of null drift, because they familiarized the animals to the environment for a substantial time period prior to recording. We refer to the second phase as a type of drift, because it happens after learning has finished and also features a gradual change in representations.

Mechanism of sparsification

What are the mechanisms that give rise to this observed sparsification? As illustrated in **Fig. 1**, solutions in the zero-loss manifold have identical loss, but might vary in some of their properties. The authors of suggest that noisy learning will slowly increase the flatness of the loss landscape in the vicinity of the solution. This can be demonstrated with a simple example. Consider a two-dimensional loss function. The function is shaped like a valley with a continuous one-dimensional zero-loss manifold at its bottom (**Fig 5A**). Crucially, the loss on this one-dimensional manifold is exactly zero, while the vicinity of the manifold becomes systematically flatter in one direction. We simulated gradient descent with added noise on this function from a random starting point (red dot). The trajectory quickly converged to the zero-loss manifold, and began a random walk on it. This walk was clearly biased towards the flatter area of the manifold, as can be seen by the spread of the trajectory. This bias could be comprehended by noting that the gradient was orthogonal to the contour lines of the loss, and therefore had a component directed towards the flat region.

In higher dimensions, flatness is captured by the eigenvalues of the Hessian of the loss. Because these eigenvalues are a collection of numbers, different scenarios could lead to minimizing different aspects of this collection. Specifically, according to [26], update noise should regularize the sum of the log of the non-zero eigenvalues while label noise should do the same for the sum of

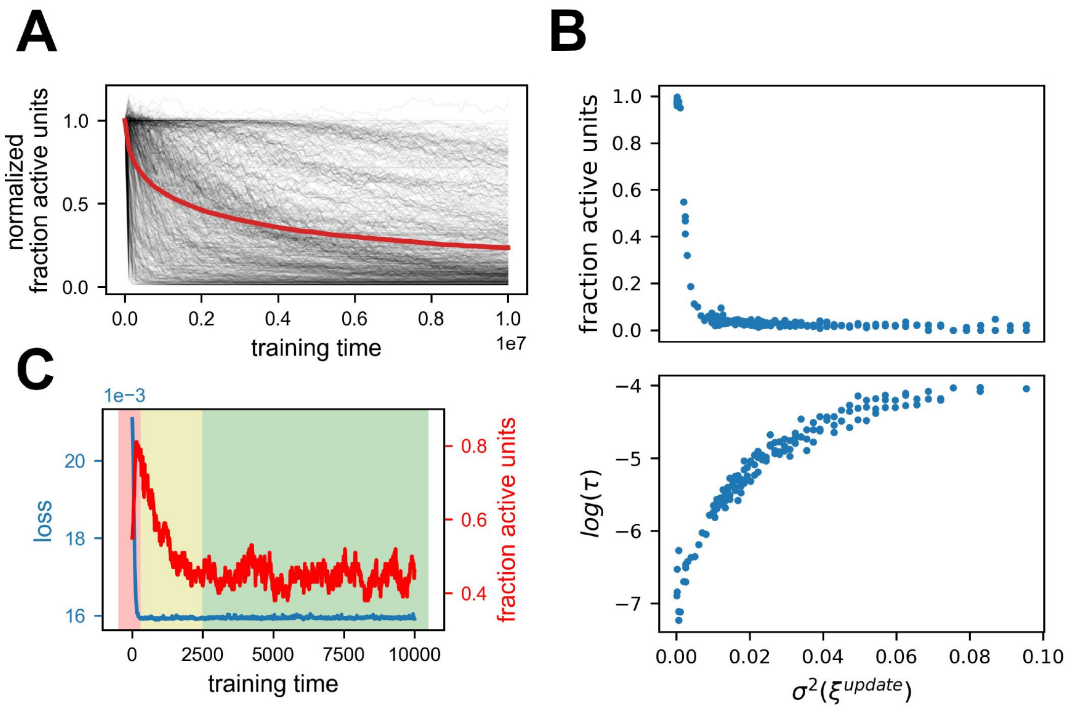


Figure 4

Generality of the results.

Summary of 616 simulations with various parameters, excluding SGD with label noise (see **Table 2**). (A) Fraction of active units normalized by the first timestep for all simulations. Red line is the mean. Note that all simulations exhibit a stochastic decrease in the fraction of active units. See **Fig S1** for further breakdown. (B) Dependence of sparseness (top) and sparsification time scale (bottom) on noise amplitude. Each point is one of 178 simulations with the same parameters except noise variance. (C) Learning a similarity matching task with Hebbian and anti-Hebbian learning using published code from [24]. Performance of the network (blue) and fraction of active units (red) as a function of training steps. Note that the loss axis does not start at zero, and the dynamic range is small. The background colors indicate which phase is dominant throughout learning (1 - red, 2 - yellow, 3 - green).

	Phase	Duration	Performance	Activity Statistics	Representations
1	learning of task	short	changing	changing	changing
2	directed drift	long	stationary	changing	changing
3	null drift	endless	stationary	stationary	changing

Table 1

The three phases of noisy learning.

Parameter	Possible values
learning algorithm	{SGD, Adam, SED}
noise type	{update, label}
number of samples	[1,20]
initialization regime	{lazy, rich}
task	{abstract predictive, random, random smoothed}
input dimension	[1,20]
output dimension	[1,20]
noise variance (label/update)	[0.1,1]/[0.01,0.1]
hidden layer size	100

Table 2

Parameter ranges for random simulations.

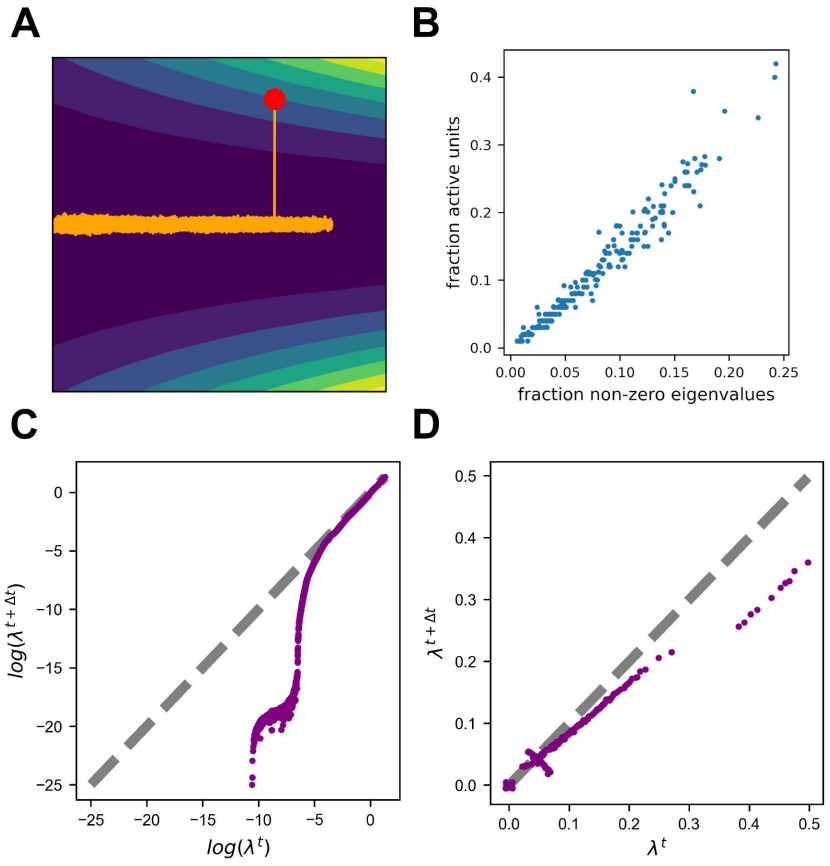


Figure 5

Noisy learning leads to a flat landscape.

(A) Gradient Descent dynamics over a two-dimensional loss function with a one-dimensional zero-loss manifold (colors from blue to yellow denote loss). Note that the loss is identically zero along the horizontal axis, but the left area is flatter. The orange trajectory begins at the red dot. Note the asymmetric extension into the left area. (B) Fraction of active units is highly correlated with the number of non-zero eigenvalues of the Hessian. (C) Update noise reduces small eigenvalues. Log of non-zero eigenvalues at two consecutive time points for learning with update noise. Note that eigenvalues do not correspond to one another when calculated at two different time points, and this plot demonstrates the change in their distribution rather than changes in eigenvalues corresponding to specific directions. The distribution of larger eigenvalues hardly changes, while the distribution of smaller eigenvalues is pushed to smaller values. (D) Label noise reduces the sum over eigenvalues. Same as (C), but for actual values instead of log.

eigenvalues. In our predictive coding example, where update noise was added, each inactivated unit translates into a set of zero-rows in the Hessian (see methods), and thus also into a set of zero-eigenvalues (**Fig 5B**). The slope of the regularizer approaches infinity as the eigenvalue approaches zero, and thus small eigenvalues are driven to zero much faster than large eigenvalues (**Fig 5C**). So in this case, update noise leads to an increase in the number of zero eigenvalues, which are manifested as a sparse solution. Another, perhaps more intuitive, way to understand these results is that units below the activation threshold are insensitive to noise perturbations. In other scenarios, in which we simulated with label noise, we indeed observed a gradual decrease in the sum of eigenvalues (**Fig 5D**). For a more intuitive demonstration of this phenomenon, see **Fig S2**.

Discussion

We showed that representational drift could arise from ongoing learning in the presence of noise, after a network has already reached good performance. We suggest that learning is divided into three overlapping phases: a fast initial phase, where good performance is achieved, a second slower phase in which *directed* drift along the low-loss manifold leads to an implicit regularization and finally, a third *undirected* phase ensues once the regularizer is minimized. In our results, the directed component was associated with sparsification of the neural code. We verified the existence of this phenomenon in experimental data from four different labs. It is important to note that sparseness was related to flatness of the loss landscape in the specific case of a single hidden layer feedforward neural network and update or label noise. For other architectures and noise types, the change in activity statistics will most likely be different and calls for further work. The CA1 region is known to have little recurrent connections, which possibly explains why these results match.

Interpreting drift as a learning process has recently been suggested by [24, 25]. Both studies focused on the final phase in which the statistics of the representations were constant. Experimentally, [7] reported a decrease in activity at the beginning of the experiment, which they suggested was correlated with some behavioral change, but we believe it could also be a result of the directed drift phase. [40] also reported a slow directed change in representation long after familiarity with the stimuli. There is another consequence of the timescale separation. Unlike in the setting of drift experiments, natural environments are never truly constant. Thus, it is possible that the second phase of learning never stops because the task is slowly changing. This would imply that the second, directed, phase may be the natural regime in which neural networks reside.

Here, we reported directed drift in the space of solutions of neural networks. This drift could be observed by examining changes to the representation of external world variables, and hence is related to the phenomenon of representational drift. Note, however, that representations are not a full description of a network's behavior [41]. The statistics of representational changes can be used as a window into changes of network dynamics and function.

The phenomenon of directed drift is very robust to various modeling choices, and also consistent with recent theoretical results [26, 27] The details of the direction of the drift, however, are dependent on specific choices. Specifically, which aspects of the Hessian are minimized during the second phase of learning, as well as the timescale of this phase, depend on the specifics of the learning rule and the noise in the system. This suggests an exciting opportunity – inferring the learning rule of a network from the statistics of representational drift.

Our explanation of drift invoked the concept of a low-loss manifold – a family of network configurations that have nearly identical performance on a task. The definition of low-loss, however, depends on the specific task and context analyzed. Challenging a system with new inputs

could dissociate two configurations that otherwise appear identical [42]. It will be interesting to explore whether various environmental perturbations could uncover the motion along the low-loss manifold in the CA1 population. An important subject relating to perturbations is that of remapping – the phenomenon in which place cells change their tuning in response to a change in the environment or change in context. One can therefore speculate that, as the network moves to flatter areas of the loss landscape, becoming more robust to noise and thus to perturbations, the probability for remapping given the same environmental change will systematically decrease. A functional interpretation of remapping as latent state (context) inference is given by [43]. The authors also summarize results from a series of morph experiments and offer a predictions similar to ours (Fig. 7 there) – that discrepancies between remapping probabilities can by explained as testing at different points of training. Beyond such abstract models, for future work, it is also possible to mechanistically model multiple environments and study remapping probabilities through them [44].

Machine learning has been suggested as a model tool for neuroscience research [45, 46, 47]. However, the implicit regularization in ML has not been studied to explain representational drift in neuroscience, and may have been done without awareness of this phenomenon. It's worth noting that this isn't a phenomenon specific to neural networks, but rather a general property of overparameterized systems that optimize a cost function. Importing insights from this domain into neuroscience shows the utility of studying general phenomena in systems that learn. For example, another complex learning system in which a similar idea has been proposed is evolution – "survival of the flattest" suggests that, under a high mutation rate, the fittest replicators are not just the ones with the highest fitness, but also with a flat fitness function which is more robust to mutations [48]. One can hope that more such insights will arise as we open our eyes.

Code availablity

The code for this project is available at *https://github.com/Aviv-Ratzon/DriftReg*

Materials and methods Predictive coding task

The agent is moving in an arena of size (L_x, L_y), with constant velocity in the y direction of V_0. The agent's heading direction is θ and it changes at every time step by $\Delta\theta \sim G(0, \sigma_\theta^2)$, the agent's visual field has an angle θ_{vis} and is represented as a vector of size L_{vis}. The texture of the walls is generated from a random Gaussian vector of size $L_{walls} = 2(L_x + L_y)L_{vis}$, smoothed with a Gaussian filter with $\sigma^2 = K_{smooth}L_{walls}$. At each time step the agent receives the visual input from the walls, determined by the intersection points of it's visual field with the walls. When the agent reaches a distance of L_yL_{buffer} from the wall, it turns to the opposite direction.

Tuning properties of units

For each unit we calculated a tuning curve. We divided the arena into 100 equal bins and computed the number of time steps in each bin and the mean unit activation. We then obtained the tuning curve by dividing the mean activity for each bin by the occupancy. We treated movement in each direction as a separate location. We calculated the spatial information (SI) of the tuning curves for each unit:

$$SI = \sum_i p_i \frac{r_i}{\bar{r}} \log_2 \frac{r_i}{\bar{r}} \tag{5}$$

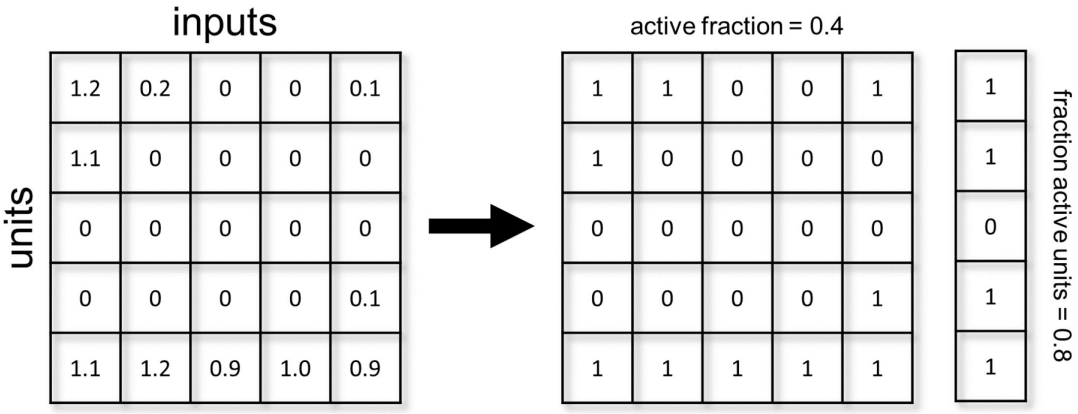
where i is the index of the bin, p_i is the probability of being in the bin, r_i is the value of the tuning curve in the bin and $\bar{r}$ is the unit's mean activity rate. Active unit was defined as a unit with non-zero activation for at least one input.

Measuring sparsity

We measure sparsity using two metrics: Active fraction, and fraction active units. These can be calculated from the activation matrix, where each row corresponds to a single unit, each column corresponds to an input and the values of the cells are the activations of a given unit for a given input. We can then binarize this matrix, giving a value of 1 to cells with non-zero activations. The active fraction is the mean over all cells of the matrix. The fraction of active units is the fraction of rows with at least one non-zero value. Note that "units" refers to hidden layer units.

Simulations

For the random simulations, we train each network for 10^7 training steps while choosing random learning algorithm and parameters. The ranges and relevant values of parameters are specified in **Table 2**. For Adam and SED there was no added noise.



Sparsification timescale

To produce **Fig 4B** bottom, we fitted an exponential curve over the curve of fraction of active units for each simulation and extracted the time constant of the exponential. To simplify this fit, we shifted each curve such that it plateaus at zero. We also clipped the length of the curves at the point in which they reach 90% of their final value to avoid fitting noise in the plateau after convergence. Note that we did this calculation to a subset of 178 simulations with the same parameters and varying noise scale. We chose this set of parameters because it exhibited a relatively "nice" exponential curve. As can be seen is **Fig 4** A, some simulations exhibit a long plateau in the fraction of active units followed by a sudden start of the sparsification process. This type of plateau followed by a phase transition was described in previous works [49, 50]. For some of the simulations finding the timescale of sparsification is not straightforward and rather noisy, thus for simplicity sake we chose to calculate it only over the mentioned subset.

Stochastic Error Descent

The equation for parameter updates under this learning rule is given by:

$$\theta_{\tau+1} = \theta_\tau - \eta(f(\theta_\tau + \xi_\tau) - f(\theta_\tau))\xi_\tau \tag{6}$$

In this learning rule, the parameters are randomly perturbed at each training step by a Gaussian noise denoted by ξ_τ and then updated in proportion to the change in loss.

Experimental data

We present here a detailed description of the analyses performed for each data set. **Table 3** summarizes the differences between them, along with an additional publication about CA1 drift [23] in which stationary statistics were reported.

Khatib et al. [22]

The results presented here were also shown in the paper, and a full description is available there. Only frames where the mice moved faster then $1\frac{cm}{s}$ were analyzed. We calculated the fraction of place cells out of all recognized cells, place cells were classified using a shuffle test. We also used the published code from [51] to verify that the increase in SI is sustained under bias corrections. Note that we treated the linear track as one-dimensional and separated the two different running directions, bins were 4*cm* in length. The metrics were averaged over cells pooled together from all animals. Data is not publicly available.

Jercog et al. [52]

The results presented here are novel. The published data features rate maps and full trajectories, without spike data. Only neurons with an average activity rate of over 0.1*Hz* were included. Because of this, we calculated only the number of active neurons each day rather than the fraction and could not classify which were place cells. We calculated the binned occupancy maps from the trajectories and used them with the published rate maps to calculate SI. Data is available at: *https://crcns.org/data-sets/hc/hc-22/about-hc-22*

Karlsson et al. [53]

The results presented here are novel. The data features spikes and trajectories. We filtered the data to include time bins when the animals moved faster than $1\frac{cm}{s}$ and only neurons from CA1. We used the published code from [51] to calculate the fraction of place cells along with SI, and verified that the increase in SI is sustained under bias corrections. For the occupancy map we treated the entire W-shaped arena as a square and and used bins of approximately 4*cm* length. This was done for the sake of simplifying the analysis, a more accurate method would be to consider separately each linear track and movement direction. Note that in this dataset animals spent varying amounts of time in two different environments. We pooled together for each animal the data of cells from either environment according to experience time in them. For example, if the animal visited environment A on day 4 for the 4th time and had 20 active cells, and visited environment B on day 6 for the 4th time and had 30 active cells, we pooled together the entire 50 cells for day 4 of this animal. Data is available at: *https://crcns.org/data-sets/hc/hc-6/about-hc-5*

Sheintuch et al. [35]

The results presented here were also shown in the paper. The data features various metrics calculated from the activity. We averaged the fraction of place cells and mean SI over animals for each day. Data is available at: *https://github.com/zivlab/cell_assemblies*

Label noise

Label noise is introduced to the loss function given by the following formula:

$$f(\theta_\tau) \;=\; \mathbb{E}_t(\hat{\mathbf{y}}_t - \mathbf{x}_{t+1} + \xi_\tau^{label})^2, \tag{7}$$

where ξ_τ^{label} is Gaussian noise.

	Khatib et al. [22]	Jercog et al. [33]	Karlsson et al. [53]	Sheintuch et al. [35]	Geva et al. [23]
familiarity	3-5 days	novel	novel	novel	6-9 days
animals	mice	mice	rats	mice	mice
# animals	8	12	9	8	8
recordings days	1 day	10 days	max. 11 days	10 days	10 days
session length	200 minutes	40 minutes	15-30 minutes	20 minutes	20 minutes
recording type	calcium imaging	electrophysiology	electrophysiology	calcium imaging	calcium imaging
arena	linear track	square or circle	W-shaped	linear or L-shaped track	linear track
activity metric	fraction of place cells decrease	number of active cells decrease	fraction of place cells decrease	fraction of place cells decrease	fraction of place cell stationary
mean SI change	increase	increase	increase	increase	stationary

Table 3

Description of experimental data sets.

Gradient descent dynamics around the zero-loss manifold

The function we used for the two-dimensional example was given by:

$$L(x,y) \;=\; (xy)^2, \tag{8}$$

which has zero loss on the x and y axes. For small enough update noise, GD will converge to the vicinity of this manifold (the axes). We consider a point on the x axis: (x_0, 0), and calculate the direction of the gradient near that point. Because we are interested in motion along the zero-loss manifold, we consider a small perturbation in the orthogonal direction (x_0, 0 + Δy) where $x_0 >> 1$ and $|\Delta y| << 1$. Any component of the gradient in the x direction will lead to motion along the manifold. The update step at this point is given by:

$$-\nabla L(x_0, 0+\Delta y) \;=\; -2x_0 \begin{pmatrix} (\Delta y)^2 \\ x_0 \Delta y \end{pmatrix}. \tag{9}$$

One can observe that the step has a large component in the y direction, quickly returning to the manifold. There is also a smaller component in the x direction, reducing the value of x. Reducing x also reduces the Hessian's eigenvalues:

$$H_L(x_0,0) = 2\begin{pmatrix} 0 & 0 \\ 0 & x_0^2 \end{pmatrix} \tag{10}$$

$$\lambda_{1,2} = \{0, x_0^2\}, v_{1,2} = \{\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}\}. \tag{11}$$

Thus, it becomes clear that the trajectory will have a bias that reduces the curvature in the y direction.

For general loss functions and various noise models, rigorous proofs can be found in [26], and a different approach can be found in [27]. Here, we will briefly outline the intuition for the general case. Consider again the update rule for GD:

$$\theta \leftarrow \theta - \eta \nabla L(\theta). \tag{12}$$

In order to understand the dynamics close to the zero-loss manifold, we consider a point θ, for which $L(\theta) = 0$ expand the loss around it:

$$L(\theta + \delta\theta) = L(\theta) + \nabla^T L(\theta)\delta\theta + \frac{1}{2}\delta\theta^T H \delta\theta. \tag{13}$$

We can then take the gradient of this expansion with respect to θ:

$$\nabla_\theta L(\theta+\delta\theta) \;=\; \nabla_\theta L(\theta) + \nabla_\theta \nabla_\theta^T L(\theta)\delta\theta + \nabla_\theta(\frac{1}{2}\delta\theta^T H \delta\theta) \tag{14}$$

$$\;=\; 0 + H\delta\theta + \nabla_\theta(\frac{1}{2}\delta\theta^T H\delta\theta). \tag{15}$$

The first term is zero, because the gradient is zero on the manifold. The second term is the largest one, as it linear in $\delta\theta$. Note that the Hessian matrix has zero eigenvalues in directions on the zero-loss manifold, and non-zero eigenvalues in other directions. Thus, the second term corresponds to projecting $\delta\theta$ in a direction that is orthogonal to the zero-loss manifold. The third term can be interpreted as the gradient of some auxiliary loss function. Thus, we expect gradient descent to minimize this new loss, which corresponds to a quadratic form with the Hessian. This is the reason for the implicit regularization along the manifold. Note that the auxiliary loss function is defined by $\delta\theta$, and thus different noise statistics will correspond, on average, to different implicit regularizations. In conclusion, the update step will have a large component that moves the parameter vector towards the zero-loss manifold, and a small component that moves the parameter vector on the manifold in a direction that minimizes some measure of the Hessian.

Hessian and sparseness

In the main text, we show that the implicit regularization of the Hessian leads to sparse representations. Here, we show this relationship for a single-hidden layer feed-forward neural network with ReLU activation and Mean Squared Error loss:

$$f(\mathbf{x}_i) = \sigma(\mathbf{x}_l\mathbf{m}^T + b)\mathbf{n}^T \tag{16}$$

The gradient and Hessian at the zero-loss manifold are given by [54]:

$$\nabla_\theta f(\mathbf{x}_i) = \begin{pmatrix} \frac{\partial f}{\partial \mathbf{m}} \\ \frac{\partial f}{\partial \mathbf{b}} \\ \frac{\partial f}{\partial \mathbf{n}} \end{pmatrix} = \begin{pmatrix} \mathbf{n} \odot \mathbb{1}(\mathbf{x}_i;\theta) \otimes \mathbf{x}_i \\ \mathbf{n} \odot \mathbb{1}(\mathbf{x}_i;\theta) \\ (\mathbf{x}_i \cdot \mathbf{n}^T + \mathbf{b}) \odot \mathbb{1}(\mathbf{x}_i;\theta) \end{pmatrix} \tag{17}$$

$$\nabla^2_\theta L(\mathbf{x};\theta) = \sum_i \nabla_\theta f(\mathbf{x}_i) \nabla_\theta f(\mathbf{x}_i)^T, \tag{18}$$

where $\mathbb{1}(\mathbf{x}_i;\theta)$ is an indicator vector denoting whether each unit is active for some input $\mathbf{x}_i$. Sparseness means that a unit has become inactive for all inputs. All the partial derivatives of input, output and bias weights associated with such a unit are zero, and thus the relevant rows of the Hessian are zero as well. Thus, every inactive unit leads to several zero eigenvalues.

Acknowledgements

We thank Ron Teichner and Kabir Dabholkar for comments on the manuscript. This research was supported by the ISRAEL SCIENCE FOUNDATION (grants Nos. 2655/18 and 2183/21 to DD, and 1442/21to OB), by the German-Israeli Foundation (GIF I-1477-421.13/2018) to DD, by a grant from the US-Israel Binational Science Foundation (NIMH-BSF CRCNS BSF:2019807, NIMH:R01 MH125544- 01 to DD), by an HFSP research grant (RGP0017/2021) to OB, A Rappaport Institute Collaborative research grant to DD, by Israel PBC-VATAT and by the Technion Center for Machine Learning and Intelligent Systems (MLIS) to DD and OB, by the Prince Center for the Aging Brain, and by a University of Michigan – Israel Partnership for Research and Education Collaborative Research stipend to DK. (data science)

Supplementary

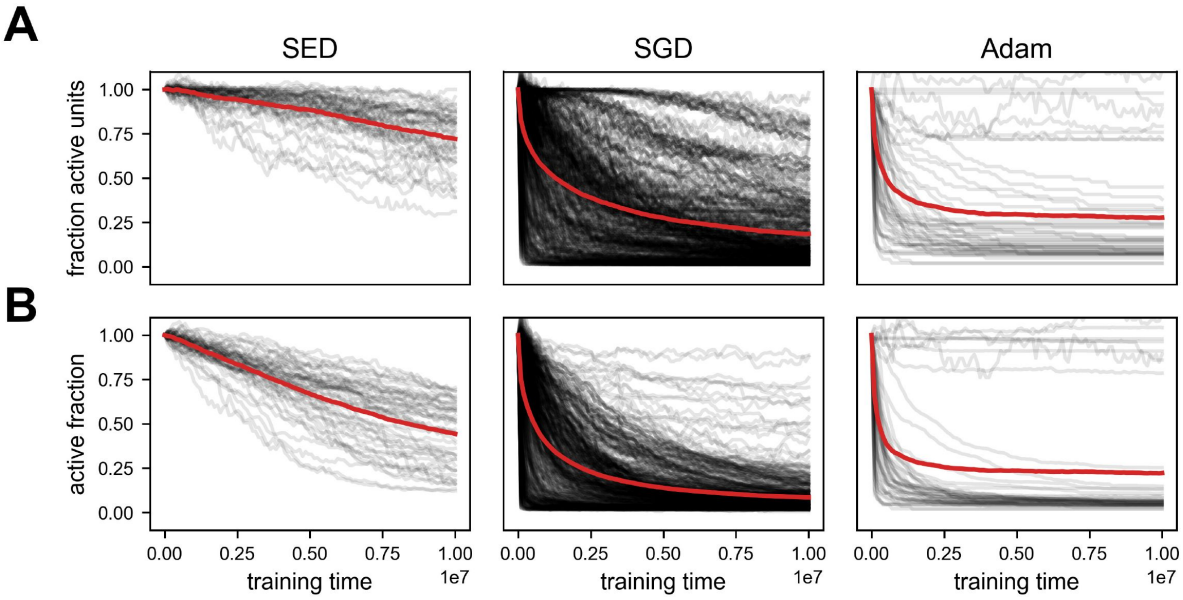


Figure S1

Noisy learning leads to spontaneous sparsification.

Summary of 516 simulations with three different learning algorithms: Stochastic error descent (SED, [38]), Stochastic gradient descent (SGD), Adam. All values are normalized to the first time step of each simulation. The red lines indicate mean over all simulations. (A) Fraction active units – number of units with any response. (B) Active fraction – overall activity across all units (see methods).

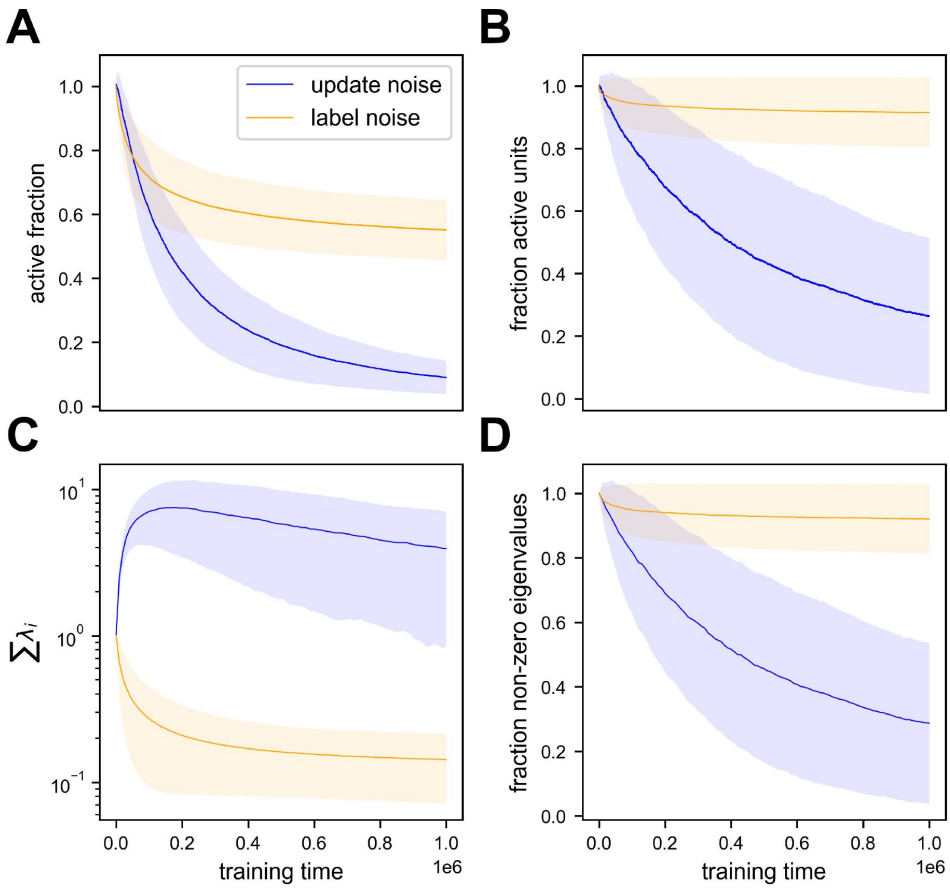


Figure S2

Label and update noise impose different regularizations over the Hessian with distinct signatures in activity statistics.

Summary of 362 simulations with either label or update noise added to SGD learning algorithm. All values are normalized to the first time step of each simulation. Lines indicate the mean of simulations and shaded regions indicate one standard deviation. Loss convergence varies between simulations, and is achieved on a scale of no more than 10^5 time steps. (A) Active fraction as a function of training time. Note this metric decreases significantly for both types of noise. (B) Fraction of active units as a function of training time. For label noise the change is much smaller. (C) Sum of the loss Hessian's eigenvalues as a function of training time. Here the difference is apparent - label noise imposes slow implicit regularization over this metric while update noise does not. (D) Fraction of non-zero eigenvalues in the loss Hessian as a function of training time. As explained in the main text, update noise imposes implicit regularization over the sum of log-eigenvalues, which manifests as a zeroing of eigenvalues over time and thus a reduction in fraction of active units.

References

[1] O'keefe John, Nadel Lynn (1979) **The hippocampus as a cognitive map** *Behavioral and Brain Sciences* **2**:487–494

[2] O'Keefe John, Dostrovsky Jonathan (1971) **The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat** *Brain research*

[3] Hubel David H, Wiesel Torsten N (1962) **Receptive fields, binocular interaction and functional architecture in the cat's visual cortex** *The Journal of physiology* **160**

[4] McNaughton Bruce L, Mizumori SJY, Barnes CA, Leonard BJ, Marquis M, Green EJ (1994) **Cortical representation of motion during unrestrained spatial navigation in the rat** *Cerebral Cortex* **4**:27–39

[5] Ziv Yaniv, Burns Laurie D, Cocker Eric D, Hamel Elizabeth O, Ghosh Kunal K, Kitch Lacey J, El Gamal Abbas, Schnitzer Mark J (2013) **Long-term dynamics of ca1 hippocampal place codes** *Nature neuroscience* **16**:264–26

[6] Driscoll Laura N., Pettit Noah L., Minderer Matthias, Chettih Selmaan N., Harvey Christopher D. (2017) **Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex** *Cell* **170**:986–999

[7] Deitch Daniel, Rubin Alon, Ziv Yaniv (2021) **Representational drift in the mouse visual cortex** *Current biology* **31**:4327–433

[8] Schoonover Carl E, Ohashi Sarah N, Axel Richard, Fink Andrew J P (2021) **Representational drift in primary olfactory cortex** *Nature* **594**

[9] Jacobson Gilad A, Rupprecht Peter, Friedrich Rainer W (2018) **Experience-dependent plasticity of odor representations in the telencephalon of zebrafish** *Current Biology* **28**:1–1

[10] Liberti William A, Schmid Tobias A, Forli Angelo, Snyder Madeleine, Yartsev Michael M (2022) **Publisher correction: A stable hippocampal code in freely flying bats** *Nature* **606**

[11] Sadeh Sadra, Clopath Claudia (2022) **Contribution of behavioural variability to representational drift** *Elife* **11**

[12] Rule Michael E, O'Leary Timothy, Harvey Christopher D (2019) **Causes and consequences of representational drift** *Curr. Opin. Neurobiol* **58**:141–147

[13] Driscoll Laura N, Duncker Lea, Harvey Christopher D (2022) **Representational drift: Emerging theories for continual learning and experimental future directions** *Current Opinion in Neurobiology* **76**

[14] Ziv Noam E, Brenner Naama (2018) **Synaptic tenacity or lack thereof: spontaneous remodeling of synapses** *Trends in neurosciences* **41**:89–9

[15] Rubin Alon, Geva Nitzan, Sheintuch Liron, Ziv Yaniv (2015) **Hippocampal ensemble dynamics timestamp events in long-term memory** *elife* **4**

[16] Aitken Kyle, Garrett Marina, Olsen Shawn, Mihalas Stefan (2022) **The geometry of representational drift in natural and artificial neural networks** *PLOS Computational Biology* **18**

[17] Kappel David, Habenschuss Stefan, Legenstein Robert, Maass Wolfgang (2015) **Network plasticity as bayesian inference** *PLoS computational biology* **11**

[18] Rokni Uri, Richardson Andrew G, Bizzi Emilio, Sebastian Seung H (2007) **Motor learning with unstable neural representations** *Neuron* **54**:653–66

[19] Susman Lee, Brenner Naama, Barak Omri (2019) **Stable memory with unstable synapses** *Nature communications* **10**

[20] Mongillo Gianluigi, Rumpel Simon, Loewenstein Yonatan (2017) **Intrinsic volatility of synaptic con-nections—a challenge to the synaptic trace theory of memory** *Current opinion in neurobiology* **46**:7–1

[21] Kossio Yaroslav Felipe Kalle, Goedeke Sven, Klos Christian, Memmesheimer Raoul-Martin (2021) **Drifting assemblies for persistent memory: Neuron transitions and unsupervised compensation** *Proceedings of the National Academy of Sciences* **118**

[22] Khatib Dorgham, Ratzon Aviv, Sellevoll Mariell, Barak Omri, Morris Genela, Derdikman Dori (2023) **Active experience, not time, determines within-day representational drift in dorsal ca1** *Neuron*

[23] Geva Nitzan, Deitch Daniel, Rubin Alon, Ziv Yaniv (2023) **Time and experience differentially affect distinct aspects of hippocampal representational drift** *Neuron*

[24] Qin Shanshan, Farashahi Shiva, Lipshutz David, Sengupta Anirvan M, Chklovskii Dmitri B, Pehlevan Cengiz (2023) **Coordinated drift of receptive fields in hebbian/anti-hebbian network models during noisy representation learning** *Nature Neuroscience* :1–1

[25] Pashakhanloo Farhad, Koulakov Alexei (2023) **Stochastic gradient descent-induced drift of representation in a two-layer neural network** *arXiv preprint*

[26] Blanc Guy, Gupta Neha, Valiant Gregory, Valiant Paul (2020) **Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process** *In Conference on learning theory* :483–513

[27] Li Zhiyuan, Wang Tianhao, Arora Sanjeev (2021) **What happens after sgd reaches zero loss?–a mathematical framework** *arXiv preprint*

[28] Yang Ning, Tang Chao, Tu Yuhai (2023) **Stochastic gradient descent introduces an effective landscapedependent regularization favoring flat solutions** *Physical Review Letters* **130**

[29] Bengio Yoshua, Lee Dong-Hyun, Bornschein Jorg, Mesnard Thomas, Lin Zhouhan (2015) **Towards biologically plausible deep learning** *arXiv preprint*

[30] Liu Yuhan Helena, Ghosh Arna, Richards Blake, Shea-Brown Eric, Lajoie Guillaume (2022) **Beyond accuracy: generalization properties of bio-plausible temporal credit assignment rules** *Advances in Neural Information Processing Systems* **35**:23077–2309

[31] Marschall Owen, Cho Kyunghyun, Savin Cristina (2020) **A unified framework of online learning algorithms for training recurrent neural networks** *The Journal of Machine Learning Research* **21**:5320–535

[32] Recanatesi Stefano, Farrell Matthew, Lajoie Guillaume, Deneve Sophie, Rigotti Mattia, Shea-Brown Eric (2021) **Predictive learning as a network mechanism for extracting low-dimensional latent space representations** *Nature Communications* **12**

[33] Jercog Pablo E, Ahmadian Yashar, Woodruff Caitlin, Deb-Sen Rajeev, Abbott Laurence F, Kandel Eric R (2019) **Heading direction with respect to a reference point modulates place-cell activity** *Nature communications* **10**

[34] Karlsson Mattias P, Frank Loren M (2008) **Network dynamics underlying the formation of sparse, informative representations in the hippocampus** *Journal of Neuroscience* **28**:14271–1428

[35] Sheintuch Liron, Geva Nitzan, Deitch Daniel, Rubin Alon, Ziv Yaniv (2023) **Organization of hippocampal ca3 into correlated cell assemblies supports a stable spatial code** *Cell Reports* **42**

[36] Kingma Diederik P, Ba Jimmy (2014) **Adam: A method for stochastic optimization** *arXiv preprint*

[37] Hinton Geoffrey, Srivastava Nitish, Swersky Kevin (2012) **Neural networks for machine learning lecture 6a overview of mini-batch gradient descent** *Cited on* **14**

[38] Cauwenberghs Gert (1992) **A fast stochastic error-descent algorithm for supervised learning and opti-mization** *Advances in neural information processing systems* **5**

[39] Mulayoff Rotem, Michaeli Tomer, Soudry Daniel (2021) **The implicit bias of minima stability: A view from function space** *Advances in Neural Information Processing Systems* **34**:17749–1776

[40] Nguyen Nghia D, Lutas Andrew, Fernando Jesseba, Vergara Josselyn, McMahon Justin, Dimidschstein Jordane, Andermann Mark L (2022) **Cortical reactivations predict future sensory responses** *bioRxiv* :2022–1

[41] Brette Romain (2019) **Is coding a relevant metaphor for the brain?** *Behavioral and Brain Sciences* **42**

[42] Turner Elia, Dabholkar Kabir V, Barak Omri (2021) **Charting and navigating the space of solutions for recurrent neural networks** *Advances in Neural Information Processing Systems* **34**:25320–2533

[43] Sanders Honi, Wilson Matthew A, Gershman Samuel J (2020) **Hippocampal remapping as hidden state inference** *Elife* **9**

[44] Low Isabel IC, Giocomo Lisa M, Williams Alex H (2023) **Remapping in a recurrent neural network model of navigation and context inference** *bioRxiv* :2023–0

[45] Richards Blake A *et al.* (2019) **A deep learning framework for neuroscience** *Nature neuroscience* **22**:1761–177

[46] Marblestone Adam H, Wayne Greg, Kording Konrad P (2016) **Toward an integration of deep learning and neuroscience** *Frontiers in computational neuroscience*

[47] Saxe Andrew, Nelli Stephanie, Summerfield Christopher (2021) **If deep learning is the answer, what is the question?** *Nature Reviews Neuroscience* **22**:55–6

[48] Codoñer Francisco M, Darós José-Antonio, Solé Ricard V, Elena Santiago F (2006) **The fittest versus the flattest: experimental confirmation of the quasispecies effect with subviral pathogens** *PLoS pathogens* **2**

[49] Saxe Andrew M, McClelland James L, Ganguli Surya (2013) **Exact solutions to the nonlinear dynamics of learning in deep linear neural networks** *arXiv preprint*

[50] Schuessler Friedrich, Mastrogiuseppe Francesca, Dubreuil Alexis, Ostojic Srdjan, Barak Omri (2020) **The interplay between randomness and structure during learning in rnns** *Advances in neural information processing systems* **33**:13352–1336

[51] Sheintuch Liron, Rubin Alon, Ziv Yaniv (2022) **Bias-free estimation of information content in temporally sparse neuronal activity** *PLoS computational biology* **18**

[52] Jercog PE, Abbott LF, Kandel ER (2019) **Hippocampal CA1 neurons recording from mice foraging in three different environments over 10 days** *CRCNS.org*

[53] Karlsson M, Carr M, Frank LM (2015) **Simultaneous extracellular recordings from hippocampal areas ca1 and ca3 (or mec and ca1) from rats performing an alternation task in two wshapped tracks that are geometrically identically but visually distinct** *CRCNS* https://doi.org/10.6080/K0NK3BZJ

[54] Nacson Mor Shpigel, Mulayoff Rotem, Ongie Greg, Michaeli Tomer, Soudry Daniel (2023) **The implicit bias of minima stability in multivariate shallow relu networks** *arXiv preprint*

Article and author information

Aviv Ratzon
Rappaport Faculty of Medicine, Technion - Israel Institute of Technology, Haifa 31096, Israel, Network Biology Research Laboratory, Technion - Israel Institute of Technology, Haifa 32000, Israel
For correspondence: aviv.ratzon@hotmail.com
ORCID iD: 0009-0000-7648-9744

Dori Derdikman
Rappaport Faculty of Medicine, Technion - Israel Institute of Technology, Haifa 31096, Israel
ORCID iD: 0000-0003-3677-6321

Omri Barak
Rappaport Faculty of Medicine, Technion - Israel Institute of Technology, Haifa 31096, Israel, Network Biology Research Laboratory, Technion - Israel Institute of Technology, Haifa 32000, Israel
ORCID iD: 0000-0002-7894-6344

Copyright

© 2023, Ratzon et al.

This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Editors

Reviewing Editor
Adrien Peyrache
McGill University, Montreal, Canada

Senior Editor
Laura Colgin
University of Texas at Austin, Austin, United States of America

Reviewer #1 (Public Review):

The authors start from the premise that neural circuits exhibit "representational drift" -- i.e., slow and spontaneous changes in neural tuning despite constant network performance. While the extent to which biological systems exhibit drift is an active area of study and debate (as the authors acknowledge), there is enough interest in this topic to justify the development of theoretical models of drift.

The contribution of this paper is to claim that drift can reflect a mixture of "directed random motion" as well as "steady state null drift." Thus far, most work within the computational neuroscience literature has focused on the latter. That is, drift is often viewed to be a harmless byproduct of continual learning under noise. In this view, drift does not affect the performance of the circuit nor does it change the nature of the network's solution or representation of the environment. The authors aim to challenge the latter viewpoint by showing that the statistics of neural representations can change (e.g. increase in sparsity) during early stages of drift. Further, they interpret this directed form of drift as "implicit regularization" on the network.

The evidence presented in favor of these claims is concise, but on balance I find their evidence persuasive, at least in artificial network models. This paper includes a brief analysis of four independent experiments in Figure 3, which corroborates the main claims of the paper. Future work should dig deeper into the experimental data to provide a finer grained characterization. For example, in addition to quantifying the overall number of active units, it would be interesting to track changes in the signal-to-noise ratio of each place field, the widths of the place fields, et cetera.

To establish the possibility of implicit regularization in artificial networks, the authors cite convincing work from the machine learning community (Blanc et al. 2020, Li et al., 2021). Here the authors make an important contribution by translating these findings into more biologically plausible models and showing that their core assumptions remain plausible. The authors also develop helpful intuition in Figure 5 by showing a minimal model that captures the essence of their result.

https://doi.org/10.7554/eLife.90069.2.sa2

Reviewer #2 (Public Review):

Summary:

In the manuscript "Representational drift as a result of implicit regularization" the authors study the phenomenon of representational drift (RD) in the context of an artificial network which is trained in a predictive coding framework. When trained on a task for spatial navigation on a linear track, they found that a stochastic gradient descent algorithm led to a fast initial convergence to spatially tuned units, but then to a second very slow, yet directed drift which sparsified the representation while increasing the spatial information. They finally show that this separation of time-scales is a robust phenomenon and occurs for a number of distinct learning rules.

This is a very clearly written and insightful paper, and I think people in the community will benefit from understanding how RD can emerge in such artificial networks. The mechanism underlying RD in these models is clearly laid out and the explanation given is convincing.

It still remains unclear how this mechanism may account for the learning of multiple environments, although this is perhaps a topic for future study. The non-stationarity of the drift in this framework would seem, at first blush, to contrast with what one sees experimentally, but the authors provide compelling evidence that there are continuous changes in network properties during learning and that stationarity may be the hallmark of overfamiliarized environments. Future experimental work may further shed light on differences in RD between novel and familiar environments.

https://doi.org/10.7554/eLife.90069.2.sa1

Reviewer #3 (Public Review):

Summary:

Single unit neural activity tuned to environmental or behavioral variables gradually changes over time. This phenomenon, called representational drift, occurs even when all external variables remain constant, and challenges the idea that stable neural activity supports the performance of well-learned behaviors. While a number of studies have described representational drift across multiple brain regions, our understanding of the underlying mechanism driving drift is limited. Ratzon et al. propose that implicit regularization - which occurs when machine learning networks continue to reconfigure after reaching an optimal solution - could provide insights into why and how drift occurs in neurons. To test this theory, Ratzon et al. trained a recurrent neural network (RNN) trained to perform the oft-utilized linear track behavioral paradigm and compare the changes in hidden layer units to those observed in hippocampal place cells recorded in awake, behaving animals.

Ratzon et al. clearly demonstrate that hidden layer units in their model undergo consistent changes even after the task is well-learned, mirroring representational drift observed in real hippocampal neurons. They show that the drift occurs across three separate measures: the active proportion of units (referred to as sparsification), spatial information of units, and correlation of spatial activity. They continue to address the conditions and parameters under which drift occurs in their model to assess the generalizability of their findings to non-spatial tasks. Last, they investigate the mechanism through which sparsification occurs, showing that flatness of the manifold near the solution can influence how the network reconfigures. The authors suggest that their findings indicate a three stage learning process: 1) fast initial learning followed by 2) directed motion along a manifold which transitions to 3) undirected motion along a manifold.

Overall, the authors' results support the main conclusion that implicit regularization in machine learning networks mirrors representational drift observed in hippocampal place cells. Their findings promise to open new fields of inquiry into the connection between

machine learning and representational drift in other, non-spatial learning paradigms, and to generate testable predictions for neural data.

Strengths:

(1) Ratzon et al. make an insightful connection between well-known phenomena in two separate fields: implicit regularization in machine learning and representational drift in the brain. They demonstrate that changes in a recurrent neural network mirror those observed in the brain, which opens a number of interesting questions for future investigation.

(2) The authors do an admirable job of writing to a large audience and make efforts to provide examples to make machine learning ideas accessible to a neuroscience audience and vice versa. This is no small feat and aids in broadening the impact of their work.

(3) This paper promises to generate testable hypotheses to examine in real neural data, e.g., that drift rate should plateau over long timescales (now testable with the ability to track single-unit neural activity across long time scales with calcium imaging and flexible silicon probes). Additionally, it provides another set of tools for the neuroscience community at large to use when analyzing the increasingly high-dimensional data sets collected today.

Weaknesses:

The revised manuscript addresses all the weaknesses outlined in my initial review. However, there is one remaining (minor) weakness regarding how "sparseness" is used and defined.

Sparseness can mean different things to different fields. For example, for engram studies, sparseness could be measured at the population level by the proportion of active cells, whereas for a physiology study, sparseness might be measured at the neuron level by the change in peak firing rate of each cell as an animal enters that cell's place field. In this manuscript, the idea of "sparseness" is introduced indirectly in the last paragraph of the introduction as "...changes in activity statistics (sparseness)...", but it is unclear from the preceding text if the referenced "activity statistics" used to define sparseness are the "fraction of active units," or their "tuning specificity," or both. While sparseness is clearly defined in the Methods section for the RNN, there is no mention of how it is defined for neural data, and spatial information is not mentioned at all. For clarity, I suggest explicitly defining sparseness for both the RNN and real neural data early in the main text, e.g. "Here, we measure sparseness in neural data by A and B, and by the analogous metric(s) of X and Y in our RNN..." This is a small but important nuance that will enhance the ease of reading for a broad neuroscience audience.

https://doi.org/10.7554/eLife.90069.2.sa0

Author response:

The following is the authors' response to the original reviews.

> ***eLife assessment***
>
> *This study presents a new and valuable theoretical account of spatial representational drift in the hippocampus. The evidence supporting the claims is convincing, with a clear and accessible explanation of the phenomenon. Overall, this study will likely attract researchers exploring learning and representation in both biological and artificial neural networks.*

We would like to ask the reviewers to consider elevating the assessment due to the following arguments. As noted in the original review, the study bridges two different fields (machine

learning and neuroscience), and does not only touch a single subfield (representational drift in neuroscience). In the revision, we also analysed data from four different labs, strengthening the evidence and the generality of the conclusions.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *The authors start from the premise that neural circuits exhibit "representational drift" -- i.e., slow and spontaneous changes in neural tuning despite constant network performance. While the extent to which biological systems exhibit drift is an active area of study and debate (as the authors acknowledge), there is enough interest in this topic to justify the development of theoretical models of drift.*
>
> *The contribution of this paper is to claim that drift can reflect a mixture of "directed random motion" as well as "steady state null drift." Thus far, most work within the computational neuroscience literature has focused on the latter. That is, drift is often viewed to be a harmless byproduct of continual learning under noise. In this view, drift does not affect the performance of the circuit nor does it change the nature of the network's solution or representation of the environment. The authors aim to challenge the latter viewpoint by showing that the statistics of neural representations can change (e.g. increase in sparsity) during early stages of drift. Further, they interpret this directed form of drift as "implicit regularization" on the network.*
>
> *The evidence presented in favor of these claims is concise. Nevertheless, on balance, I find their evidence persuasive on a theoretical level -- i.e., I am convinced that implicit regularization of noisy learning rules is a feature of most artificial network models. This paper does not seem to make strong claims about real biological systems. The authors do cite circumstantial experimental evidence in line with the expectations of their model (Khatib et al. 2022), but those experimental data are not carefully and quantitatively related to the authors' model.*

We thank the reviewer for pushing us to present stronger experimental evidence. We now analysed data from four different labs. Two of those are novel analyses of existing data (Karlsson et al, Jercog et al). All datasets show the same trend - increasing sparsity and increasing information per cell. We think that the results, presented in the new figure 3, allow us to make a stronger claim on real biological systems.

> *To establish the possibility of implicit regularization in artificial networks, the authors cite convincing work from the machine-learning community (Blanc et al. 2020, Li et al., 2021). Here the authors make an important contribution by translating these findings into more biologically plausible models and showing that their core assumptions remain plausible. The authors also develop helpful intuition in Figure 4 by showing a minimal model that captures the essence of their result.*

We are glad that these translation efforts are appreciated.

> *In Figure 2, the authors show a convincing example of the gradual sparsification of tuning curves during the early stages of drift in a model of 1D navigation. However, the evidence presented in Figure 3 could be improved. In particular, 3A shows a histogram displaying the fraction of active units over 1117 simulations. Although there is a spike near zero, a sizeable portion of simulations have greater than 60% active units at the end of the training, and critically the authors do not characterize the time course of the active fraction for every network, so it is difficult to evaluate their claim that "all [networks] demonstrated... [a] phase of directed random motion with the low-loss space."*

> *It would be useful to revise the manuscript to unpack these results more carefully. For example, a histogram of log(tau) computed in panel B on a subset of simulations may be more informative than the current histogram in panel A.*

The previous figure 3A was indeed confusing. In particular, it lumped together many simulations without proper curation. We redid this figure (now Figure 4), and added supplementary figures (Figures S1, S2) to better explain our results. It is now clear that the simulations with a large number of active units were either due to non-convergence, slow timescale of sparsification or simulations featuring label noise in which the fraction of active units is less affected. Regarding the log(tau) calculation, while it could indeed be an informative plot, it could not be calculated in a simple manner for all simulations. This is because learning curves are not always exponential, but sometimes feature initial plateaus (see also Saxe et al 2013, Schuessler et al 2020). We added a more detailed explanation of this limitation in the methods section, and we believe the current figure exemplifies the effect in a satisfactory manner.

> ***Reviewer #2 (Public Review):***
>
> *Summary:*
>
> *In the manuscript "Representational drift as a result of implicit regularization" the authors study the phenomenon of representational drift (RD) in the context of an artificial network that is trained in a predictive coding framework. When trained on a task for spatial navigation on a linear track, they found that a stochastic gradient descent algorithm led to a fast initial convergence to spatially tuned units, but then to a second very slow, yet directed drift which sparsified the representation while increasing the spatial information. They finally show that this separation of timescales is a robust phenomenon and occurs for a number of distinct learning rules.*
>
> *Strengths:*
>
> *This is a very clearly written and insightful paper, and I think people in the community will benefit from understanding how RD can emerge in such artificial networks. The mechanism underlying RD in these models is clearly laid out and the explanation given is convincing.*

We thank the reviewer for the support.

> *Weaknesses:*
>
> *It is unclear how this mechanism may account for the learning of multiple environments.*

There are two facets to the topic of multiple environments. First, are the results of the current paper relevant when there are multiple environments? Second, what is the interaction between brain mechanisms of dealing with multiple environments and the results of the current paper?

We believe the answer to the first question is positive. The near-orthogonality of representations between environments implies that changes in one can happen without changes in the other. This is evident, for instance, in Khatib et al and Geva et al - in both cases, drift seems to happen independently in two environments, even though they are visited intermittently and are visually similar.

The second question is a fascinating one, and we are planning to pursue it in future work. While the exact way in which the brain achieves this near-independence is an open question, remapping is one possible window into this process.

We extended the discussion to make these points clear.

> *The process of RD through this mechanism also appears highly non-stationary, in contrast to what is seen in familiar environments in the hippocampus, for example.*

The non-stationarity noted by the reviewer is indeed a major feature of our observations, and is indeed linked to familiarity. We divide learning into three phases (now more clearly stated in Table 1 and Figure 4C). The first, rapid phase, consists of improvement of performance - corresponding to initial familiarity with the environment. The third phase, often reported in the literature of representational drift, is indeed stationary and obtained after prolonged familiarity. Our work focuses on the second phase, which is not as immediate as the first one, and can take several days. We note in the discussion that experiments which include a long familiarization process can miss this phase (see also Table 3). Furthermore, we speculate that real life is less stationary than a lab environment, and this second phase might actually be more relevant there.

> ***Reviewer #3 (Public Review):***
>
> *Summary:*
>
> *Single-unit neural activity tuned to environmental or behavioral variables gradually changes over time. This phenomenon, called representational drift, occurs even when all external variables remain constant, and challenges the idea that stable neural activity supports the performance of well-learned behaviors. While a number of studies have described representational drift across multiple brain regions, our understanding of the underlying mechanism driving drift is limited. Ratzon et al. propose that implicit regularization - which occurs when machine learning networks continue to reconfigure after reaching an optimal solution - could provide insights into why and how drift occurs in neurons. To test this theory, Ratzon et al. trained a Feedforward Network trained to perform the oft-utilized linear track behavioral paradigm and compare the changes in hidden layer units to those observed in hippocampal place cells recorded in awake, behaving animals.*
>
> *Ratzon et al. clearly demonstrate that hidden layer units in their model undergo consistent changes even after the task is well-learned, mirroring representational drift observed in real hippocampal neurons. They show that the drift occurs across three separate measures: the active proportion of units (referred to as sparsification), spatial information of units, and correlation of spatial activity. They continue to address the conditions and parameters under which drift occurs in their model to assess the generalizability of their findings.*
>
> *However, the generalizability results are presented primarily in written form: additional figures are warranted to aid in reproducibility.*

We added figures, and a Github with all the code to allow full reproducibility.

> *Last, they investigate the mechanism through which sparsification occurs, showing that the flatness of the manifold near the solution can influence how the network reconfigures. The authors suggest that their findings indicate a three-stage learning process: 1) fast initial learning followed by 2) directed motion along a manifold which transitions to 3) undirected motion along a manifold.*
>
> *Overall, the authors' results support the main conclusion that implicit regularization in machine learning networks mirrors representational drift observed in hippocampal place cells.*

We thank the reviewer for this summary.

> *However, additional figures/analyses are needed to clearly demonstrate how different parameters used in their model qualitatively and quantitatively influence drift.*

We now provide additional figures regarding parameters (Figures S1, S2).

> *Finally, the authors need to clearly identify how their data supports the three-stage learning model they suggest.*
>
> *Their findings promise to open new fields of inquiry into the connection between machine learning and representational drift and generate testable predictions for neural data.*
>
> *Strengths:*
>
> *(1) Ratzon et al. make an insightful connection between well-known phenomena in two separate fields: implicit regularization in machine learning and representational drift in the brain. They demonstrate that changes in a recurrent neural network mirror those observed in the brain, which opens a number of interesting questions for future investigation.*
>
> *(2) The authors do an admirable job of writing to a large audience and make efforts to provide examples to make machine learning ideas accessible to a neuroscience audience and vice versa. This is no small feat and aids in broadening the impact of their work.*
>
> *(3) This paper promises to generate testable hypotheses to examine in real neural data, e.g., that drift rate should plateau over long timescales (now testable with the ability to track single-unit neural activity across long time scales with calcium imaging and flexible silicon probes). Additionally, it provides another set of tools for the neuroscience community at large to use when analyzing the increasingly high-dimensional data sets collected today.*

We thank the reviewer for these comments. Regarding the hypotheses, these are partially confirmed in the new analyses we provide of data from multiple labs (new Figure 3 and Table 3) - indicating that prolonged exposure to the environment leads to more stationarity.

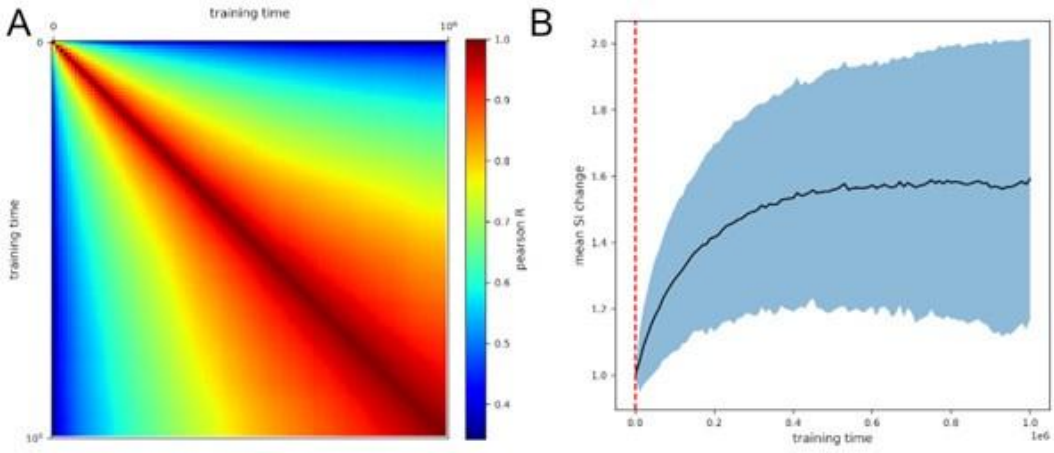
> *Weaknesses:*
>
> *(1) Neural representational drift and directed/undirected random walks along a manifold in ML are well described. However, outside of the first section of the main text, the analysis focuses primarily on the connection between manifold exploration and sparsification without addressing the other two drift metrics: spatial information and place field correlations. It is therefore unclear if the results from Figures 3 and 4 are specific to sparseness or extend to the other two metrics. For example, are these other metrics of drift also insensitive to most of the Feedforward Network parameters as shown in Figure 3 and the related text? These concerns could be addressed with panels analogous to Figures 3a-c and 4b for the other metrics and will increase the reproducibility of this work.*

We note that the results from figures 3 and 4 (original manuscript) are based on abstract tasks, while in figure 2 there is a contextual notion of spatial position. Spatial position metrics are not applicable to the abstract tasks as they are simple random mapping of inputs, and there isn't necessarily an underlying latent variable such as position. This transition between task types is better explained in the text now. In essence the spatial information and place

field correlation changes are simply signatures of the movements in parameter space. In the abstract tasks their change becomes trivial, as the spatial information becomes strongly correlated with sparsity and place fields are simply the activity vectors of units. These are guaranteed to change as long as there are changes in the activity statistics. We present here the calculation of these metrics averaged over simulations for completeness.

Author response image 1.

PV correlation between training time points averaged over 362 simulations. (B) Mean SI of units normalized to first time step, averaged over 362 simulations. Red line shows the average time point of loss convergence, the shaded area represents one standard deviation.



> *(2) Many caveats/exceptions to the generality of findings are mentioned only in the main text without any supporting figures, e.g., "For label noise, the dynamics were qualitatively different, the fraction of active units did not reduce, but the activity of the units did sparsify" (lines 116-117). Supporting figures are warranted to illustrate which findings are "qualitatively different" from the main model, which are not different from the main model, and which of the many parameters mentioned are important for reproducing the findings.*

We now added figures (S1, S2) that show this exactly. We also added a github to allow full reproduction.

> *(3) Key details of the model used by the authors are not listed in the methods. While they are mentioned in reference 30 (Recanatesi et al., 2021), they need to be explicitly defined in the methods section to ensure future reproducibility.*

The details of the simulation are detailed in the methods sections. We also added a github to allow full reproducibility.

> *(4) How different states of drift correspond to the three learning stages outlined by the authors is unclear. Specifically, it is not clear where the second stage ends, and the third stage begins, either in real neural data or in the figures. This is compounded by the fact that the third stage - of undirected, random manifold exploration - is only discussed in relation to the introductory Figure 1 and is never connected to the neural network data or actual brain data presented by the authors. Are both stages meant to represent drift? Or is only the second stage meant to mirror drift, while undirected random motion along a manifold is a prediction that could be tested in real neural data? Identifying where each stage occurs in Figures 2C and E, for example, would clearly illustrate which*

> *attributes of drift in hidden layer neurons and real hippocampal neurons correspond to each stage.*

Thanks for this comment, which urged us to better explain these concepts.

The different processes (reduction in loss, reduction in Hessian) happen in parallel with different timescales. Thus, there are no sharp transitions between the phases. This is now explained in the text in relation to figure 4C, where the approximate boundaries are depicted.

The term drift is often used to denote a change in representation without a change in behavior. In this sense, both the second and third phases correspond to drift. Only the third stage is stationary. This is now emphasized in the text and in the new Table 1. Regarding experimental data, apart from the new figure 3 with four datasets, we also summarize in Table 3 the relation between duration of familiarity and stationarity of the data.

> ***Recommendations for the authors:***
>
> *The reviewers have raised several concerns. They concur that the authors should address the specific points below to enhance the manuscript.*
>
> *(1) The three different phases of learning should be clearly delineated, along with how they are determined. It remains unclear in which exact phase the drift is observed.*

This is now clearly explained in the new Table 1 and Figure 4C. Note that the different processes (reduction in loss, reduction in Hessian) happen in parallel with different timescales. Thus, there are no sharp transitions between the phases. This is now explained in the text in relation to figure 4C, where the approximate boundaries are depicted.

The term drift is often used to denote a change in representation without a change in behavior. In this sense, both the second and third phases correspond to drift. Only the third stage is stationary. This is now emphasized in the text and in the new Table 1. Regarding experimental data, apart from the new figure 3 with four datasets, we also summarize in Table 3 the relation between duration of familiarity and stationarity of the data.

> *(2) The term "sparsification" of unit activity is not fully clear. Its meaning should be more explicitly explained, especially since, in the simulations, a significant number of units appear to remain active (Fig. 3A).*

We now define precisely the two measures we use - Active Fraction, and Fraction Active Units. There is a new section with an accompanying figure in the Methods section. As Figure S2 shows, the noise statistics (label noise vs. update noise) differentially affects these two measures.

> *(3) While the study primarily focuses on one aspect of representational drift-the proportion of active units-it should also explore other features traditionally associated with representational drift, such as spatial information and the correlation between place fields.*

This absence of features is related to the abstract nature of some of the tasks simulated in our paper. In our original submission the transition between a predictive coding task to more abstract tasks was not clearly explained, creating some confusion regarding the measured metrics. We now clarified the motivation for this transition.

Both the initial simulation and the new experimental data analysis include spatial information (Figures 2,3). The following simulations (Figure 4) with many parameter choices use more abstract tasks, for which the notion of correlation between place cells and spatial

information loses its meaning as there is no spatial ordering of the inputs, and every input is encountered only once. Spatial information becomes strongly correlated with the inverse of the active fraction metric. The correlation between place cells is also directly linked to increase in sparseness for these tasks.

> *(4) There should be a clearer illustration of how labeling noise influences learning dynamics and sparsification.*

This was indeed confusing in the original submission. We removed the simulations with label noise from Figure 4, and added a supplementary figure (S2) illustrating the different effects of label noise.

> *(5) The representational drift observed in this study's simulations appears to be nonstationary, which differs from in vivo reports. The reasons for this discrepancy should be clarified.*

We added experimental results from three additional labs demonstrating a change in activity statistics (i.e. increase in spatial information and increase in sparseness) over a long period of time. We suggest that such a change long after the environment is already familiar is an indication for the second phase, and stress that this change seems to saturate at some point, and that most drift papers start collecting data after this saturation, hence this effect was missed in previous in vivo reports. Furthermore, these effects are become more abundant with the advent on new calcium imaging methods, as the older electrophysiological regording methods did not usually allow recording of large amounts of cells for long periods of time. The new Table 3 surveys several experimental papers, emphasizing the degree of familiarity with the environment.

> *(6) A distinctive feature of the hippocampus is its ability to learn different spatial representations for various environments. The study does not test representational drift in this context, a topic of significant interest to the community. Whether the authors choose to delve into this is up to them, but it should at least be discussed more comprehensively, as it's only briefly touched upon in the current manuscript version.*

There are two facets to the topic of multiple environments. First, are the results of the current paper relevant when there are multiple environments? Second, what is the interaction between brain mechanisms of dealing with multiple environments and the results of the current paper?

We believe the answer to the first question is positive. The near-orthogonality of representations between environments implies that changes in one can happen without changes in the other. This is evident, for instance, in Khatib et al and Geva et al - in both cases, drift seems to happen independently in two environments, even though they are visited intermittently and are visually similar.

The second question is a fascinating one, and we are planning to pursue it in future work. While the exact way in which the brain achieves this near-independence is an open question, remapping is one possible window into this process.

We extended the discussion to make these points clear.

> *(7) The methods section should offer more details about the neural nets employed in the study. The manuscript should be explicit about the terms "hidden layer", "units", and "neurons", ensuring they are defined clearly and not used interchangeably..*

We changed the usage of these terms to be more coherent and made our code publicly available. Specifically, “units” refer to artificial networks and “neurons” to biological ones.

> *In addition, each reviewer has raised both major and minor concerns. These are listed below and should be addressed where possible.*
>
> ***Reviewer #1 (Recommendations For The Authors):***
>
> *I recommend that the authors edit the text to soften their claims. For example:*
>
> *In the abstract "To uncover the underlying mechanism, we..." could be changed to "To investigate, we..."*

Agree. Done

> *On line 21, "Specifically, recent studies showed that..." could be changed to "Specifically, recent studies suggest that..."*

Agree. Done

> *On line 100, "All cases" should probably be softened to "Most cases" or more details should be added to Figure 3 to support the claim that every simulation truly had a phase of directed random motion.*

The text was changed in accordance with the reviewer’s suggestion. In addition, the figure was changed and only includes simulations in which we expected unit sparsity to arise (without label noise). We also added explanations and supplementary figures for label noise.

> *Unless I missed something obvious, there is no new experimental data analysis reported in the paper. Thus, line 159 of the discussion, "a phenomenon we also observed in experimental data" should be changed to "a phenomenon that recently reported in experimental data."*

We thank the reviewer for drawing our attention to this. We now analyzed data from three other labs, two of which are novel analyses on existing data. All four datasets show the same trends of sparseness with increasing spatial information. The new Figure 3 and text now describe this.

> *On line 179 of the Discussion, "a family of network configurations that have identical performance..." could be softened to "nearly identical performance." It would be possible for networks to have minuscule differences in performance that are not detected due to stochastic batch effects or limits on machine precision.*

The text was changed in accordance with the reviewer’s suggestion.

> *Other minor comments:*
>
> *Citation 44 is missing the conference venue, please check all citations are formatted properly.*

Corrected.

> *In the discussion on line 184, the connection to remapping was confusing to me, particularly because the cited reference (Sanders et al. 2020) is more of a conceptual model than an artificial network model that could be adapted to the setting of noisy*

> *learning considered in this paper. How would an RNN model of remapping (e.g. Low et al. 2023; Remapping in a recurrent neural network model of navigation and context inference) be expected to behave during the sparsifying portion of drift?*

We now clarified this section. The conceptual model of Sanders et al includes a specific prediction (Figure 7 there) which is very similar to ours - a systematic change in robustness depending on duration of training. Regarding the Low et al model, using such mechanistic models is an exciting avenue for future research.

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *I only have two major questions.*
>
> *(1) Learning multiple representations: Memory systems in the brain typically must store many distinct memories. Certainly, the hippocampus, where RD is prominent, is involved in the ongoing storage of episodic memories. But even in the idealized case of just two spatial memories, for example, two distinct linear tracks, how would this learning process look? Would there be any interference between the two learning processes or would they be largely independent? Is the separation of time scales robust to the number of representations stored? I understand that to answer this question fully probably requires a research effort that goes well beyond the current study, but perhaps an example could be shown with two environments. At the very least the authors could express their thoughts on the matter.*

There are two facets to the topic of multiple environments. First, are the results of the current paper relevant when there are multiple environments? Second, what is the interaction between brain mechanisms of dealing with multiple environments and the results of the current paper?

We believe the answer to the first question is positive. The near-orthogonality of representations between environments implies that changes in one can happen without changes in the other. This is evident, for instance, in Khatib et al and Geva et al - in both cases, drift seems to happen independently in two environments, even though they are visited intermittently and are visually similar.

The second question is a fascinating one, and we are planning to pursue it in future work. While the exact way in which the brain achieves this near-independence is an open question, remapping is one possible window into this process.

We extended the discussion to make these points clear.

> *(2) Directed drift versus stationarity: I could not help but notice that the RD illustrated in Fig.2D is not stationary in nature, i.e. the upper right and lower left panels are quite different. This appears to contrast with findings in the hippocampus, for example, Fig.3e-g in (Ziv et al, 2013). Perhaps it is obvious that a directed process will not be stationary, but the authors note that there is a third phase of steady-state null drift. Is the RD seen there stationary? Basically, I wonder if the process the authors are studying is relevant only as a novel environment becomes familiar, or if it is also applicable to RD in an already familiar environment. Please discuss the issue of stationarity in this context.*

The non-stationarity noted by the reviewer is indeed a major feature of our observations, and is indeed linked to familiarity. We divide learning into three phases (now more clearly stated in Table 1 and Figure 4C). The first, rapid, phase consists of improvement of performance - corresponding to initial familiarity with the environment. The third phase, often reported in the literature of representational drift, is indeed stationary and obtained after prolonged familiarity. Our work focuses on the second phase, which is not as immediate as the first one,

and can take several days. We note in the discussion that experiments which include a long familiarization process can miss this phase (see also Table 3). Furthermore, we speculate that real life is less stationary than a lab environment, and this second phase might actually be more relevant there.

> ***Reviewer #3 (Recommendations For The Authors):***
>
> *Most of my general recommendations are outlined in the public review. A large portion of my comments regards increasing clarity and explicitly defining many of the terms used which may require generating more figures (to better illustrate the generality of findings) or modifying existing figures (e.g., to show how/where the three stages of learning map onto the authors' data).*
>
> *Sparsification is not clearly defined in the main text. As I read it, sparsification is meant to refer to the activity of neurons, but this needs to be clearly defined. For example, lines 262-263 in the methods define "sparseness" by the number of active units, but lines 116-117 state: "For label noise, the dynamics were qualitatively different, the fraction of active units did not reduce, but the activity of the units did sparsify." If the fraction of active units (defined as "sparseness") did not change, what does it mean that the activity of the units "sparsified"? If the authors mean that the spatial activity patterns of hidden units became more sharply tuned, this should be clearly stated.*

We now defined precisely the two measures we use - Active Fraction, and Fraction Active Units. There is a new section with an accompanying figure in the Methods section. As Figure S2 shows, the noise statistics (label noise vs. update noise) differentially affects these two measures.

> *Likewise, it is unclear which of the features the authors outlined - spatial information, active proportion of units, and spatial correlation - are meant to represent drift. The authors should clearly delineate which of these three metrics they mean to delineate drift in the main text rather than leave it to the reader to infer. While all three are mentioned early on in the text (Figure 2), the authors focus more on sparseness in the last half of the text, making it unclear if it is just sparseness that the authors mean to represent drift or the other metrics as well.*

The main focus of our paper is on the non-stationarity of drift. Namely that features (such as these three) systematically change in a directed manner as part of the drift process. This is in The new analyses of experimental data show sparseness and spatial information.

The focus on sparseness in the second half of the paper is because we move to more abstract These are also easy to study in the more abstract tasks in the second part of the paper. In our original submission the transition between a predictive coding task to more abstract tasks was not clearly explained, creating some confusion regarding the measured metrics. We now clarified the motivation for this transition.

> *It is not clear if a change in the number of active units alone constitutes "drift", especially since Geva et al. (2023) recently showed that both changes in firing rate AND place field location drive drift, and that the passage of time drives changes in activity rate (or # cells active).*

Our work did not deal with purely time-dependent drift, but rather focused on experience-dependence. Furthermore, Geva et al study the stationary phase of drift, where we do not expect a systematic change in the total number of cells active. They report changes in the average firing rate of active cells in this phase, as a function of time - which does not contradict our findings.

> *"hidden layer", "units", and "neurons" seem to be used interchangeably in the text (e.g., line 81-85). However, this is confusing in several places, in particular in lines 83-85 where "neurons" is used twice. The first usage appears to refer to the rate maps of the hidden layer units simulated by the authors, while the second "neurons" appears to refer to real data from Ziv 2013 (ref 5). The authors should make it explicit whether they are referring to hidden layer units or actual neurons to avoid reader confusion.*

We changed the usage of these terms to be more coherent. Specifically, “units” refer to artificial networks and “neurons” to biological ones.

> *The authors should clearly illustrate which parts of their findings support their three-phase learning theory. For example, does 2E illustrate these phases, with the first tenth of training time points illustrating the early phase, time 0.1-0.4 illustrating the intermediate phase, and 0.4-1 illustrating the last phase? Additionally, they should clarify whether the second and third stages are meant to represent drift, or is it only the second stage of directed manifold exploration that is considered to represent drift? This is unclear from the main text.*

The different processes (reduction in loss, reduction in Hessian) happen in parallel with different timescales. Thus, there are no sharp transitions between the phases. This is now explained in the text in relation to figure 4C, where the approximate boundaries are depicted.

The term drift is often used to denote a change in representation without a change in behavior. In this sense, both the second and third phases correspond to drift. Only the third stage is stationary. This is now emphasized in the text and in the new Table 1. Regarding experimental data, apart from the new figure 3 with four datasets, we also summarize in Table 3 the relation between duration of familiarity and stationarity of the data.

> *Line 45 - It appears that the acronym ML is not defined above here anywhere.*

Added.

> *Line 71: the ReLU function should be defined in the text, e.g., sigma(x) = x if x > 0 else 0.*

Added.

> *106-107: Figures (or supplemental figures) to demonstrate how most parameters do not influence sparsification dynamics are warranted. As written, it is unclear what "most parameters" mean - all but noise scale. What about the learning rule? Are there any interactions between parameters?*

We now removed the label noise from Figure 4, and added two supplementary figures to clearly explain the effect of parameters. Figure 4 itself was also redone to clarify this issue.

> *2F middle: should "change" be omitted for SI?*

The panel was replaced by a new one in Figure 3.

> *116-119: A figure showing how results differ for label noise is warranted.*

This is now done in Figure S1, S2.

> *124: typo, The -> the*

Corrected.

> *127-129: This conclusion statement is the first place in the text where the three stages are explicitly outlined. There does not appear to be any support or further explanation of these stages in the text above.*

We now explain this earlier at the end of the Introduction section, along with the new Table 1 and marking on Figure 4C.

> *132-133 seems to be more of a statement and less of a prediction or conclusion - do the authors mean "the flatness of the loss landscape in the vicinity of the solution predicts the rate of sparsification?"*

We thank the reviewer for this observation. The sentence was rephrased:

Old: As illustrated in Fig. 1, different solutions in the zero-loss manifold might vary in some of their properties. The specific property suggested from theory is the flatness of the loss landscape in the vicinity of the solution.

New: As illustrated in Fig. 1, solutions in the zero-loss manifold have identical loss, but might vary in some of their properties. The authors of [26] suggest that noisy learning will slowly increase the flatness of the loss landscape in the vicinity of the solution.

> *135: typo, it's -> its*

Corrected.

> *Line 135-136 "Crucially, the loss on the 136 entire manifold is exactly zero..." This appears to contradict the Figure 4A legend - the loss appears to be very high near the top and bottom edges of the manifold in 4A. Do the authors mean that the loss along the horizontal axis of the manifold is zero?*

The reviewer is correct. The manifold mentioned in the sentence is indeed the horizontal axis. We changed the text and the figure to make it clearer.

> *Equation 6: This does not appear to agree with equation 2 - should there be an E_t term for an expectation function?*

Corrected.

> *Line 262-263: "Sparseness means that a unit has become inactive for all inputs." This should also be stated explicitly as the definition of sparseness/sparsification in the main text.*

We now define precisely the two measures we use - Active Fraction, and Fraction Active Units. There is a new section with an accompanying figure in the Methods section. As Figure S2 shows, the noise statistics (label noise vs. update noise) differentially affects these two measures.