Representational drift as a window into neural and behavioural plasticity

Charles Micou[1] and Timothy O'Leary[1,2]

[1]Department of Engineering, University of Cambridge, United Kingdom
[2]Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology Graduate University, Onna, 904-0495, Japan

22nd May 2023

Abstract

Large scale recordings of neural activity over days and weeks have revealed that neural representations of familiar tasks, precepts, and actions continually evolve without obvious changes in behaviour. We hypothesize that this steady drift in neural activity and accompanying physiological changes is due in part to the continuous application of a learning rule at the cellular and population level. Explicit predictions of this drift can be found in neural network models that use iterative learning to optimise weights. Drift therefore provides a measurable signal that can reveal systems-level properties of biological plasticity mechanisms, such as their precision and effective learning rates.

Introduction

> "I am tomorrow, or some future day, what I establish today. I am today what I established yesterday or some previous day."
>
> James Joyce

Animals benefit from having a continually updating model of the world. Accordingly, our brains never stop learning. This may be in the form of subtle perceptual or motor learning that occurs when we repeatedly practise a task; it may correspond to continual laying down of episodic memories; or it may arise during moments of surprise or sudden insight [1].

On the other hand, and for sound practical reasons, experimental neuroscience often separates "learning" from "doing." Laboratory learning paradigms study changes in behaviour, physiology, or both, as animals learn new tasks. A traditional paradigm breaks this process into a baseline (prior to learning), a learning or induction phase (where the animal or neural circuit is subjected to a learning paradigm), and a steady state (where behavioural performance plateaus). But from the perspective of the underlying circuitry in the nervous system there is no necessary distinction between these phases. The gradual tuning and consolidation

of learned information, the replay and parsing of past events, and the process of introspection all contribute to behaviourally relevant changes [2]. Even if potent modulatory signals convey novelty or reward in a specific time window, the synaptic and intrinsic mechanisms that form the substrate for learning must be in a state of perpetual readiness to adapt, and may need to do so long after a salient stimulus has passed [3, 4]. This raises the question of whether it is advantageous or even practical for the nervous system to switch plasticity on and off in its more malleable circuits. A more realistic scenario may entail continual change, with long term consistency enforced by dynamic interactions, both internally and with the external world.

In parallel, the theoretical study of learning algorithms developed its own split between "learning" and "doing". Measures of loss or task error bottom-out after sufficient iterations of a learning rule, at which point a model is deemed "trained" and fit for evaluation on test data. Understandably, the dynamics of the internal state of the model after the loss function stops improving have been of relatively little interest: ways to optimise learning and robustness have been the priority. Nonetheless, analyses of artificial learning systems have revealed forms of drift as an inherent feature of ongoing learning [5, 6]. We want to point out the possibility of using drift as a surrogate measure of the collective properties of learning rules, potentially allowing key characteristics of learning rules to be inferred. For this to apply to living nervous systems, there needs to be measurable drift.

Numerous recent studies indeed report that physiological correlates of behaviour - concretely, the activity patterns in populations of neurons - seem to continually drift over time, even when behaviour is apparently unchanging [7, 8, 9, 10, 11, 12]. In tandem, and likely contributing to this drift, synaptic turnover in many parts of the brain remains significant outside periods of marked learning and brain development [13, 14, 15, 16]. Much of this work only became possible with recording technology that can track populations of individual cells over the course of days and weeks using optical or electrical means [8, 10, 11].

In this review, we explore the idea that representational drift corresponds to the endgame of plasticity mechanisms operating under the necessary constraint of limited information about the external world. We do not rule out the possibility that some component of drift is due to cell biological processes that are unrelated to learning, such as the stochastic turnover of synapses. Similarly, we do not argue against some stochastic component of drift having an algorithmic purpose, such as providing a means of exploration [17] or regularisation during learning [18]. Instead, we want to point out that even if the nervous system were composed of perfectly reliable components, existing theory predicts a never-ending flow in the configuration space of any adaptive neural circuit. Measurement of this flow, in turn, yields information about the learning efficiency and mechanistic properties of plasticity mechanisms, and the effective degrees of freedom in a neural representation. Representational drift may therefore be a quantifiable signal that can allow us to study learning and plasticity at the population level.

Before outlining these ideas and the recent work that inspires them, we want to flag some terminology. Neuroscience is an interdisciplinary field, and key phrases are used with varying

degrees of consistency. We are not attempting to universally define terms here, rather define key terms as we will use them here to minimise misinterpretation. By "representation", we mean any spatiotemporal pattern of neural activity that correlates with defined actions or sensations. Questions of the causal link between the two remain on the table because they are the subject of ongoing research in all cases. Secondly, we are deliberately loose with the term "learning": we take this to mean any systematic changes in neural circuitry that are caused by, and ultimately influence, interactions with the world. This includes latent learning, which may reconfigure how information is represented in the brain without necessarily producing obvious or immediately detectable changes in behaviour. Finally, the reader will notice that there is sometimes no clear line between "plasticity" and "learning" in our discussion: we believe that that this topic necessarily blurs such a line.

Empirical statistics of drift

The feature that distinguishes representational drift from the umbrella term "neural variability" is that there appears to be no steady-state representation about which the moment-to-moment representation randomly fluctuates: the further apart in time any two representations, the more different they are from each other [10, 19, 11, 8, 7]. In other words, drift does not appear to be mean-reverting, especially not on timescales for which other brain circuitry may simply average-out fluctuations.

We must stress that any claim about the statistics of drift is subject to experimental limitations, and we must acknowledge the heterogeneity of observations across experimental conditions. Even with current recording technology, it is remarkably difficult to record individual cells over timescales of weeks. Reported timescales of drift range from a single experimental session [12] to several days or weeks [10, 19, 11], or across these timescales [8, 7]. Drift is neither a universal nor a homogeneous phenomenon. Drift rates vary across brain regions [20], being slow to negligible in some sensory and motor areas [19, 7] and higher in structures such as the hippocampus [11]. Indeed, the fact that there is circuit-dependent variation in drift rate suggests that it is an interpretable signal - a point we return to in later sections.

For the purpose of our discussion we will assume that experimental measurements can be taken at face value, and that the interesting component of drift cumulatively perturbs representations over timescales relevant for retaining learned behaviour (i.e. days or weeks). The inferences we will draw from this assumption will establish working hypotheses that directly relate drift to learning mechanisms. Questions surrounding the validity of these hypotheses should also motivate future studies to extend the timescales over which neural activity and behaviour can be precisely tracked.

Three principal sources of drift

The elephant in the room for the experimental characterisation of drift is that unmeasured behavioural changes may account for a large component of changes in neural activity

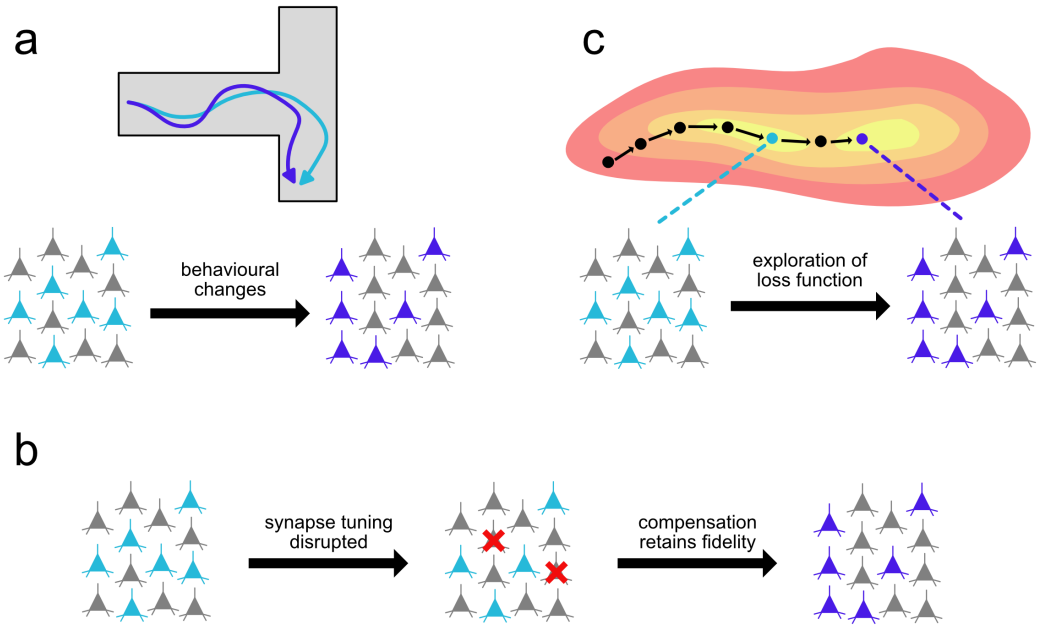

Figure 1: There are mutually compatible explanations for the observation of representational drift in neural populations. (a) Behavioural changes, albeit in ways that do not affect task performance, would still drive changes in the behaviour of the activity of the population. For example: an animal may take a slightly different trajectory through a maze, but reach the goal in roughly the same time. (b) Out-of-task factors, such as the disruption of synapse tuning due to stochastic turnover, may damage the fidelity of the representation and must then subsequently compensated for by recruiting other neurons in the population. (c) The learning algorithm implemented by the population continues exploring the solution space rather than stopping, even though it passes through perfectly adequate minima.

[21, 22, 23, 24] (Figure 1.a). Even in brain regions largely devoted to sensory processing, we find abundant examples of motor behaviour modulating representation [25, 26]. Even if it were realistic to measure and account for every aspect of an animal's externally observable behaviour, representations are also subject to change based on 'internal' states of the animal, such as their mood, tiredness, or engagement with the task [27, 28, 29]. It is therefore immensely challenging to design experiments that can completely eliminate confounds associated with behaviour.

Other, less philosophically challenging factors may account for drift. Ongoing changes in population activity may be due to the cellular and molecular makeup of synapses and neurons (Figure 1.b). Biochemical signalling is stochastic in nature and some component of turnover may be intrinsic to these mechanisms [15, 16, 30]. Synaptic turnover takes place in the absence of the learning of new behaviours [13], and some component of this turnover is even independent of neural activity [31, 32]. One may also consider issues with recording methodology, though the studies we cite go to huge lengths to account for such contributions and we are confident that not all measurements can be put down to recording technology.

Finally, and most relevant to our main point: there may be a component of drift that is attributable to the learning mechanism itself (Figure 1.c). We will expand on this possibility in later sections, but for now, we can summarise three mutually compatible sources of drift:

1. changes in behaviour that do not affect performance of a task but nevertheless impact neural activity
2. synaptic and physiological change caused by out-of-task factors that must be compensated for in order to preserve a neural representation
3. drift that is somehow intrinsic to the learning and plasticity rules implemented by a neural population

In the remainder of the paper we accept that all three sources can account for drift to some extent. We focus on the second and third sources as ways of relating drift to plasticity mechanisms because these relate most directly to theories and models of plasticity. However, we will also speculate that non-stationary changes behaviour - perhaps driven by drift in neural activity - may have a role in exploration and learning.

Drift as a consequence of ongoing learning and systematic plasticity mechanisms

We expect the space of feasible neural representations to be degenerate. That is, we expect many configurations of a neural population to represent a task or environment at comparable levels of fidelity. This is implied both by the existence of drift and by the co-existence of distinct representations of the same environment within the same time-frame [33, 34]. We picture these equivalently-performing representations as different points in a configuration space that map to locations in the basins and valleys of some loss landscape (Figure 2.a).

Recent attempts to explain drift using artificial neural networks assume synaptic noise or noisy data [18, 35], hypothesising that this will drive a random walk through degenerate regions of the loss landscape. For a high dimensional state space, a random walk deviates very rapidly from its starting point and is vanishingly unlikely to return to an earlier point on the trajectory [36, 37]. Such a process will have non mean-reverting statistics, and the effective dimensionality of the process can be inferred from experimental data, yielding clues about the intrinsic degrees of freedom in a representation. Rule and colleagues [9] performed such analyses on data from parietal cortex, finding that drift resembled a random walk in a space of far fewer dimensions than the number of recorded neurons. This estimated dimension was consistent across animals, and thus may be interpreted as a measure of the spare degrees of freedom in the neural representation. Before expanding on why this is interesting, we must stress that inferring statistical properties demands caution. Even a mean-reverting process (such as an Ornstein-Uhlenbeck process) in a high dimensional space and/or with a long autocorrelation time may have an approximately linear trend on experimental timescales, and may thus appear to be non-mean-reverting. Caveats considered, observations of how drift is constrained can reveal hallmarks of specific underlying neural system properties:

- circuit connectivity constraints (e.g. common inputs could reduce degrees of freedom)
- physiological mechanisms (e.g. (anti) Hebbian or homeostatic plasticity can cause neurons to (de) correlate and/or distribute activity patterns over time [51, 55])
- task complexity (e.g. a more complex task would consume more degrees of freedom, reducing drift dimensionality)
- task representation efficiency (e.g. more efficient codes consume fewer degrees of freedom, increasing drift dimensionality).

Finer-grained empirical properties of drift can further eliminate or support models of underlying mechanisms. For example, Qin and colleagues [55] showed that continual action of anti-Hebbian plasticity results in *coordinated drift* of units that encode different features. Rule & O'Leary showed that continual homeostatic normalisation in the presence of drift can cause downstream tuning curves to jump preference abruptly [51]. Such qualitative properties of models can make satisfyingly granular predictions about what we should see in data, bringing into reach a number of tantalising open questions.

How else might drift inform us about learning and plasticity mechanisms? Perhaps surprisingly, drift can exist during the steady state of a learning algorithm, even when data and weight updates are *not* corrupted with noise. Further back in the literature on artificial neural networks, drift was anticipated as a side effect of learning continuously on the same data distribution. Writing in 1985, McClelland and Rumelhart [6] comment on this effect in the context of an associative neural network being sequentially trained on nominal 'dog' patterns:

> The [network] acquires a set of weights which is *continually buffeted about* by the latest dog exemplar, but which captures the prototype dog quite well.

Thus, in continual learning, drift results from an optimisation process acting on recent exemplar subsets of available data as a practical means of assimilating a full dataset. This situation encompasses a class of algorithms, including stochastic gradient descent (SGD), where artificial neural networks are trained sequentially on exemplars that are typically drawn randomly. Interestingly, recently developed continual learning algorithms explicitly penalise drift to enforce stable readouts [38] or to reduce the likelihood of damaging previously learned representations [39, 40].

Learning-induced drift is symptomatic of any learning system with limited resources that prevent it from accessing and operating on all training data at once. Biologically, we would expect this to be a generic situation: it necessarily holds when learning occurs online, as neural circuitry only has access to current or recent information. It will also hold for any plausible scheme in which neural populations "replay" past experiences for the purposes of consolidation and generalisation [41, 42].

The reader may be suspicious that the kind of learning-induced drift we are describing here should die away over time as learning asymptotes, and thus not really be relevant to ongoing drift as it is measured in learned tasks. However, this intuition is wrong. In a sophisticated analysis, Chaudhari and Soatto [5] prove that continually applying SGD results in drift at steady state, even for noise-free training data. We might initially believe that randomly drawn batches are responsible for a random-walk-like exploration of the loss function about some minimum, but this is not the case. The state of the network will follow a structured flow through the loss landscape. The rate of this flow depends on the learning rate (which we can liken to the plasticity of a system) and the batch size (which we can liken to the variety of data presented to the system). Very recent work by Pashakhanloo and Koulakov [43] has directly explored the continuous application of SGD as an analogue to drift and predicted that frequently presented stimuli will generate slower rates of drift.

These analyses show the potential to directly relate the statistics of drift to algorithmic parameters of the learning algorithms implemented in the brain. For these reasons, we contend that representational drift can be used to quantify key properties of biological learning rules without necessarily requiring complex interventions.

Manipulating drift rates

As we begin to better understand the constraints on biological learning rules, we gain further opportunities to use drift to probe those rules. Consider the credit assignment problem: updates across a population cannot realistically be calculated globally and subsequently disseminated to each synapse. This issue is extensively explored in [45] and [46], and the current consensus is that while an implementation of backpropagation (and thus exact gradient descent) is unlikely in the brain, methods for approximating gradient descent are possible using local learning rules. Recent work has shown that such approximations introduce non-negligible biases in a weight update that cause them to deviate from pure gradient descent [47, 48]. These biases are systematic and would contribute to drift. They are thus experi-

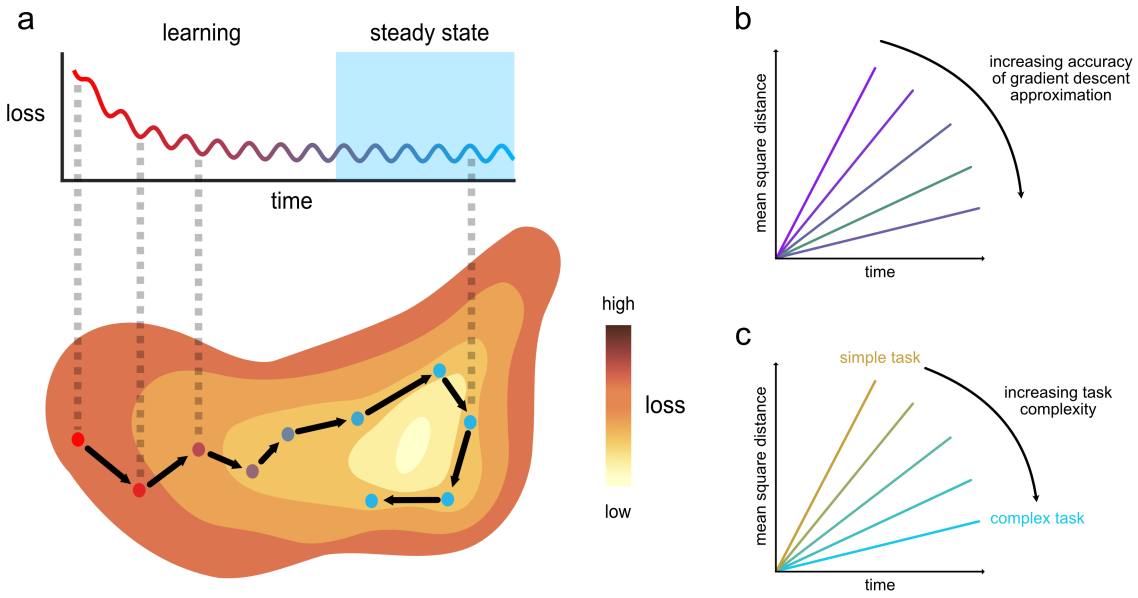

Figure 2: (a) An example trajectory through a loss landscape. After sufficient learning, the system reaches an approximate steady state. Repeated application of the learning rule, under either perturbation of the system or through the necessity to process input data continually, will continue to drive the trajectory through the loss landscape. Adapted from [44]. (b) Within the steady state region, the mean-squared distance between two samples in configuration space will increase as those samples are taken further apart in time. The rate of this drift should be higher in the cases where the trajectory through the loss landscape is governed by an update rule that is a less accurate approximation for gradient descent. (c) In the steady state, a more complex task has relatively fewer free dimensions in which to drift, and under a diffusive model of drift would therefore drift more slowly than simple task counterpart.

mentally quantifiable in principle (Figure 2.b). Recent contributions to theory show that the quality of gradient-descent approximation can be inferred by manipulating steady state synaptic turnover physiologically [44]. This work shows that if learning-related plasticity signals are blocked, synaptic turnover should decrease by a factor that scales inversely with the closeness of the approximation to gradient-descent being implemented in the circuit. It is of course extremely challenging to block plasticity cleanly, and it is more tractable to block plasticity processes than the cellular correlates of learning in an intact animal. The few studies that have attempted this show results consistent with theoretical predictions [30] - see [16] for a review.

A more experimentally tractable manipulation is the enrichment of an environment: providing more cues, objects, and concepts to be represented. When observing animal behaviour, studies find that when the degrees of freedom of animal motion exceed the dimensionality of the task, only the dimensions essential to task success are kept constrained [49, 50]. We suggest that the degenerate solution space of neural representations is similarly 'lazy': enriching an environment or scaling the number of contingencies an animal needs to learn will constrain the solution space further. Modelling work has provided a basis both for how multiple brain regions can reliably exchange information despite changing representations and how drift driven by external factors may be compensated for [9, 51, 52, 53, 54, 55]. Besides the nature of the learning rules, these models predict drift to be governed by the constraints

on the space of 'good enough' representations: a more loosely constrained population should drift faster. Thus a change to the task that consumes degrees of freedom without engaging different learning mechanisms might manifest as a change in the rate of drift (Figure 2.c). Very recent experimental work hints at such a relationship.

Khatib et al. [12] measured rates of drift in the representations of environments in CA1 hippocampus over the course of a single experimental session lasting several hours. They exposed animals to two environments over the same time period, enforcing the animal to spend substantially more time in one environment than the other. They found that the environment that the animal is exposed to more often is associated with a higher rate of drift. This suggests that additional exposure to the environment induces further iterations of the learning rule, resulting in greater drift. On the other hand, Schoonover et al. [10] measured rates of drift in the representations of odour in the olfactory cortex over the course of several weeks. They report that frequent exposure to the same odour *reduces* the drift associated with that odour's representation–seemingly the opposite of the Khatib et al. result. This is intriguing: perhaps this indicates that odour exposure, even to the same odour with no structured task, corresponds to an environmental enrichment. Regardless, these results indicate that drift rates can and do vary as a function of an animal's experience. Future experimental work could attempt to explicitly test the effect of task complexity on drift. Future theory work will help identify how drift is shaped by underlying learning rules and the neural codes they act upon.

Non-drifting circuits provide important controls and counterfactuals

We have focused attention on neural circuits that remain plastic and exhibit some degree of representational drift during nominally stable behaviour. But are there examples where stable behaviour seems to be accompanied by extremely stable neural activity?

Birdsong in zebra finches crystallises during development: once a juvenile learns its song from an adult tutor, the same base song will be repeated for life [56, 57]. This is effectively a task in which learning truly has a "before" and "after", and is doubly interesting for us because the muscle contractions required to repeat the same song are so precise that behavioural variability is necessarily very low.

Lesion experiments suggest that HVC (proper name) is the region of songbird cortex responsible for storing a lifelong representation of the learned song [58, 59]. Recent 2-photon Calcium imaging of HVC revealed stable representations with no drift over the course of several weeks [60]. This strongly implies that random, unavoidable component of drift due to synaptic turnover can be very low when there is pressure to make it so: if the brain needs an immutable circuit, it can make one.

The possibility that birdsong circuits, and likely other brain circuits, do not show representational drift strengthens the case for using drift to study plasticity: the nervous system

can stabilise neural representations when needed, suggesting that drift is not merely a result of noisy, unreliable biology. Insofar as this is the case, drift must instead have a largely systematic component, which we argue gives a window into systematic plasticity mechanisms.

Does representational drift drive behavioural drift?

A recent and comprehensive review of representational drift [61] highlights emerging ideas for how one might attempt to experimentally manipulate rates of drift without modification of the environment. The use of optogenetics to suppress individual cells within a population may increase the rate of synaptic turnover. Conversely, the use of brain-machine interfaces requiring a fixed readout may stabilise a population. These manipulations offer potential insight into the mechanisms driving drift, but they may fail to decouple representational drift from behaviour completely.

Ultimately, the purpose of neural activity is to drive behaviour. While in principle it is possible that neural representations can freely explore a space that generates equivalent behaviours, it seems an unlikely that behaviour would remain identical, even with stringent task requirements. We should therefore expect some component of drift to leak into behavioural changes, subtle or otherwise. For us, this introduces a tantalising possibility: if drift truly results in a non-stationary exploration of allowable circuit configurations, then this might eventually generate non-stationary behavioural variability. Such variability could result in gradual changes in habits, ticks, and reflexes. More broadly, this variability is potentially a useful substrate for exploring behavioural strategies [62]. In this sense, even if behavioural variability explains most of drift, we will be left with an arguably bigger question: whether behaviour and neural activity unavoidably drift together, and whether or not this is an essential part of an animal's wider ethology and adaptive capacity.

Outlook

Representational drift is a measurable signal that varies across brain regions and experimental tasks. It may exhibit qualitatively distinct statistical features depending on its underlying causes. Measurable signals give us something to work with: in the same way that systems neuroscience was revolutionised by shifting perspectives from single neuron responses to populations, drift can give us a systems-level readout of the collective action of plasticity mechanisms. However, without a quantitative framework for linking drift to hypothesised mechanisms, the phenomenon of drift is largely uninterpretable and relatively easy to dismiss. Theoretical models are crucial for providing an interpretation of drift and relating it in non-obvious ways to systems-level mechanisms of learning and memory.

Acknowledgements

This work was supported by ERC grant 716643 FLEXNEURO and was conducted while visiting the Okinawa Institute of Science and Technology (OIST) through the Theoretical

306 Sciences Visiting Program (TSVP). The authors acknowledge important ongoing discussions
307 on this topic with colleagues, especially Chris Harvey, Michael Rule, Kristine Heiney, Mónika
308 Józsa, Andrew Fink, Carl Schoonover, Claudia Clopath, Denise Cai, Alon Rubin, Yaniv Ziv,
309 Dhruva Raman, Mark Schnitzer, Kazumasa Tanaka and Laura Driscoll.

Significant Papers

1. ** Chaudhari and Soatto [5]. The authors describe the behaviour of trajectories across the loss landscape of deep networks trained using SGD, including the surprising result that these are not random walks once converged.
2. ** Sadeh and Clopath [23]. The authors draw attention to the pitfalls for drift measurements introduced by behavioural variability, focusing on the potential for misidentifying drift in populations or overestimating its extent.
3. ** Raman and O'Leary [44]. The authors show that when synaptic fluctuations are balanced by continual learning, the optimal magnitude of the learning component is typically smaller than the fluctuating component. The ratio of the components can be used to infer the quality of the plasticity rule, such as how well it approximated gradient descent.
4. ** Schoonover et al. [10]. One of the first studies to manipulate representational drift and to measure it using optimised electrophysiological methods.
5. * Katlowitz et al. [60]. The authors demonstrate a stable representation over the course of weeks in a neural population with no further need to learn representations.
6. * Khatib et al. [12] (preprint). The authors design and present new experimental data from a protocol designed to disentangle the effects of experience and time on drift, with implications for drift's direct involvement in the learning process.
7. * Lillicrap et al. [45] This review paper explores the capacity for a biological networks to implement gradient descent techniques.
8. * Pettit et al. [27]. This paper presents evidence that internal states that are not directly measurable, such as task engagement, modulate the representation of environments within the Hippocampus.

References

[1] John Kounios and Mark Beeman. The cognitive neuroscience of insight. *Annual review of psychology*, 65:71–93, 2014.

[2] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016.

[3] Richard Stuart Sutton. *Temporal credit assignment in reinforcement learning*. University of Massachusetts Amherst, 1984.

[4] Ullrich Wagner, Steffen Gais, Hilde Haider, Rolf Verleger, and Jan Born. Sleep inspires insight. *Nature*, 427(6972):352–355, 2004.

[5] Pratik Chaudhari and Stefano Soatto. Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks. In *IEEE Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, February 2018.

[6] JL McClelland and DE Rumelhart. Distributed memory and the representation of general and specific information. *Journal of experimental psychology. General*, 114(2):159–197, 1985.

[7] Daniel Deitch, Alon Rubin, and Yaniv Ziv. Representational drift in the mouse visual cortex. *Current Biology*, 31(19):4327–4339.e6, October 2021.

[8] Laura N. Driscoll, Noah L. Pettit, Matthias Minderer, Selmaan N. Chettih, and Christopher D. Harvey. Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex. *Cell*, 170(5):986–999.e16, August 2017.

[9] Michael E Rule, Adrianna R Loback, Dhruva V Raman, Laura N Driscoll, Christopher D Harvey, and Timothy O'Leary. Stable task information from an unstable neural population. *eLife*, 9:e51121, July 2020.

[10] Carl E. Schoonover, Sarah N. Ohashi, Richard Axel, and Andrew J. P. Fink. Representational drift in primary olfactory cortex. *Nature*, 594(7864):541–546, June 2021.

[11] Yaniv Ziv, Laurie D. Burns, Eric D. Cocker, Elizabeth O. Hamel, Kunal K. Ghosh, Lacey J. Kitch, Abbas El Gamal, and Mark J. Schnitzer. Long-term dynamics of CA1 hippocampal place codes. *Nature Neuroscience*, 16(3):264–266, March 2013.

[12] Dorgham Khatib, Aviv Ratzon, Mariell Sellevoll, Omri Barak, Genela Morris, and Dori Derdikman. Experience, not time, determines representational drift in the hippocampus. *bioRxiv*, 2022.

[13] Alessio Attardo, James E Fitzgerald, and Mark J Schnitzer. Impermanence of dendritic spines in live adult ca1 hippocampus. *Nature*, 523(7562):592–596, 2015.

[14] Gianluigi Mongillo, Simon Rumpel, and Yonatan Loewenstein. Intrinsic volatility of synaptic connections—a challenge to the synaptic trace theory of memory. *Current opinion in neurobiology*, 46:7–13, 2017.

[15] Yonatan Loewenstein, Uri Yanover, and Simon Rumpel. Predicting the dynamics of network connectivity in the neocortex. *Journal of Neuroscience*, 35(36):12535–12544, 2015.

[16] Noam E Ziv and Naama Brenner. Synaptic tenacity or lack thereof: spontaneous remodeling of synapses. *Trends in neurosciences*, 41(2):89–99, 2018.

[17] David Kappel, Robert Legenstein, Stefan Habenschuss, Michael Hsieh, and Wolfgang Maass. A dynamic connectome supports the emergence of stable computational function of neural circuits through reward-based learning. *eneuro*, 5(2), 2018.

[18] Kyle Aitken, Marina Garrett, Shawn Olsen, and Stefan Mihalas. The geometry of representational drift in natural and artificial neural networks. *PLOS Computational Biology*, 18(11):e1010716, November 2022.

[19] Tyler D. Marks and Michael J. Goard. Stimulus-dependent representational drift in primary visual cortex. *Nature Communications*, 12(1):5169, August 2021.

[20] Sadegh Ebrahimi, Jérôme Lecoq, Oleg Rumyantsev, Tugce Tasci, Yanping Zhang, Cristina Irimia, Jane Li, Surya Ganguli, and Mark J Schnitzer. Emergent reliability in sensory cortical coding and inter-area communication. *Nature*, 605(7911):713–721, 2022.

[21] William A. Liberti, Tobias A. Schmid, Angelo Forli, Madeleine Snyder, and Michael M. Yartsev. A stable hippocampal code in freely flying bats. *Nature*, 604(7904):98–103, April 2022.

[22] Kristopher T. Jensen, Naama Kadmon Harpaz, Ashesh K. Dhawale, Steffen B. E. Wolff, and Bence P. Ölveczky. Long-term stability of single neuron activity in the motor system. *Nature Neuroscience*, 25(12):1664–1674, December 2022.

[23] Sadra Sadeh and Claudia Clopath. Contribution of behavioural variability to representational drift. *eLife*, 11:e77907, August 2022.

[24] Simon Musall, Matthew T. Kaufman, Ashley L. Juavinett, Steven Gluf, and Anne K. Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, 22(10):1677–1686, October 2019. Number: 10 Publisher: Nature Publishing Group.

[25] Cristopher M. Niell and Michael P. Stryker. Modulation of Visual Responses by Behavioral State in Mouse Visual Cortex. *Neuron*, 65(4):472–479, February 2010.

[26] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Charu Bai Reddy, Matteo Carandini, and Kenneth D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437):eaav7893, April 2019.

[27] Noah L. Pettit, Xintong C. Yuan, and Christopher D. Harvey. Hippocampal place codes are gated by behavioral engagement. *Nature Neuroscience*, 25(5):561–566, May 2022.

[28] Pamela J. Kennedy and Matthew L. Shapiro. Motivational states activate distinct hippocampal representations to guide goal-directed behaviors. *Proceedings of the National Academy of Sciences*, 106(26):10805–10810, June 2009.

[29] Clifford G Kentros, Naveen T Agnihotri, Samantha Streater, Robert D Hawkins, and Eric R Kandel. Increased Attention to Spatial Context Increases Both Place Field Stability and Spatial Memory. *Neuron*, 42(2):283–295, April 2004.

[30] Amir Minerbi, Roni Kahana, Larissa Goldfeld, Maya Kaufman, Shimon Marom, and Noam E Ziv. Long-term relationships between synaptic tenacity, synaptic remodeling, and network activity. *PLoS biology*, 7(6):e1000136, 2009.

[31] Akira Nagaoka, Hiroaki Takehara, Akiko Hayashi-Takagi, Jun Noguchi, Kazuhiko Ishii, Fukutoshi Shirai, Sho Yagishita, Takanori Akagi, Takanori Ichiki, and Haruo Kasai. Abnormal intrinsic dynamics of dendritic spines in a fragile X syndrome mouse model in vivo. *Scientific Reports*, 6(1):26651, May 2016. Number: 1 Publisher: Nature Publishing Group.

[32] Dylan P. Quinn, Annette Kolar, Sydney A. Harris, Michael Wigerius, James P. Fawcett, and Stefan R. Krueger. The Stability of Glutamatergic Synapses Is Independent of Activity Level, but Predicted by Synapse Size. *Frontiers in Cellular Neuroscience*, 13, 2019.

[33] Liron Sheintuch, Nitzan Geva, Hadas Baumer, Yoav Rechavi, Alon Rubin, and Yaniv Ziv. Multiple Maps of the Same Spatial Context Can Stably Coexist in the Mouse Hippocampus. *Current Biology*, 30(8):1467–1476.e6, April 2020.

[34] Carol A. Barnes, Matthew S. Suster, Jiemin Shen, and Bruce L. McNaughton. Multistability of cognitive maps in the hippocampus of old rats. *Nature*, 388(6639):272–275, July 1997.

[35] Shanshan Qin, Shiva Farashahi, David Lipshutz, Anirvan M. Sengupta, Dmitri B. Chklovskii, and Cengiz Pehlevan. Coordinated drift of receptive fields in Hebbian/anti-Hebbian network models during noisy representation learning. *Nature Neuroscience*, 26(2):339–349, February 2023.

[36] Georg Pólya. Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz. *Mathematische Annalen*, 84(1):149–160, March 1921.

[37] C. Domb. On multiple returns in the random-walk problem. *Mathematical Proceedings of the Cambridge Philosophical Society*, 50(4):586–591, October 1954.

[38] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2L: Contrastive Continual Learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 9516–9525, 2021.

[39] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. Publisher: Proceedings of the National Academy of Sciences.

[40] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995. PMLR, 06–11 Aug 2017.

[41] David J. Foster. Replay Comes of Age. *Annual Review of Neuroscience*, 40(1):581–602, 2017.

[42] Laure Buhry, Amir H. Azizi, and Sen Cheng. Reactivation, Replay, and Preplay: How It Might All Fit Together. *Neural Plasticity*, 2011:203462, 2011.

[43] Farhad Pashakhanloo and Alexei Koulakov. Stochastic Gradient Descent-induced drift of representation in a two-layer neural network, February 2023. arXiv:2302.02563.

[44] Dhruva V Raman and Timothy O'Leary. Optimal plasticity for memory maintenance during ongoing synaptic change. *eLife*, 10:e62912, September 2021.

[45] Timothy P. Lillicrap, Adam Santoro, Luke Marris, Colin J. Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, 2020.

[46] James C. R. Whittington and Rafal Bogacz. Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences*, 23(3):235–250, March 2019.

[47] Dhruva Venkita Raman, Adriana Perez Rotondo, and Timothy O'Leary. Fundamental bounds on learning performance in neural circuits. *Proceedings of the National Academy of Sciences*, 116(21):10537–10546, 2019.

[48] Adriana Perez Rotondo, Dhruva V Raman, and Timothy O'Leary. How cerebellar architecture facilitates rapid online learning. *bioRxiv*, pages 2022–10, 2022.

[49] Stephen H. Scott. The computational and neural basis of voluntary motor control and planning. *Trends in Cognitive Sciences*, 16(11):541–549, November 2012.

[50] Daniel McNamee and Daniel M. Wolpert. Internal Models in Biological Control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):339–364, 2019.

[51] Michael E Rule and Timothy O'Leary. Self-healing codes: How stable neural populations can track continually reconfiguring neural representations. *Proceedings of the National Academy of Sciences*, 119(7):e2106692119, 2022.

[52] Daniel Acker, Suzanne Paradis, and Paul Miller. Stable memory and computation in randomly rewiring neural networks. *Journal of Neurophysiology*, 122(1):66–80, July 2019.

[53] Michael Jan Fauth and Mark CW Van Rossum. Self-organized reactivation maintains and reinforces memories despite synaptic turnover. *Elife*, 8:e43717, 2019.

[54] Yaroslav Felipe Kalle Kossio, Sven Goedeke, Christian Klos, and Raoul-Martin Memmesheimer. Drifting assemblies for persistent memory: Neuron transitions and unsupervised compensation. *Proceedings of the National Academy of Sciences*, 118(46):e2023832118, 2021.

[55] Shanshan Qin, Shiva Farashahi, David Lipshutz, Anirvan M Sengupta, Dmitri B Chklovskii, and Cengiz Pehlevan. Coordinated drift of receptive fields in hebbian/anti-hebbian network models during noisy representation learning. *Nature Neuroscience*, pages 1–11, 2023.

[56] Fernando Nottebohm. Auditory Experience and Song Development in the Chaffinch Fringilla Coelebs. *Ibis*, 110(4):549–568, 1968.

[57] Anthony Leonardo and Masakazu Konishi. Decrystallization of adult birdsong by perturbation of auditory feedback. *Nature*, 399(6735):466–470, June 1999.

[58] W. Hamish Mehaffey and Allison J. Doupe. Naturalistic stimulation drives opposing heterosynaptic plasticity at two inputs to songbird cortex. *Nature Neuroscience*, 18(9):1272–1280, September 2015.

[59] Dmitriy Aronov, Aaron S. Andalman, and Michale S. Fee. A Specialized Forebrain Circuit for Vocal Babbling in the Juvenile Songbird. *Science*, 320(5876):630–634, May 2008.

[60] Kalman A. Katlowitz, Michel A. Picardo, and Michael A. Long. Stable Sequential Activity Underlying the Maintenance of a Precisely Executed Skilled Behavior. *Neuron*, 98(6):1133–1140.e3, June 2018.

[61] Laura N. Driscoll, Lea Duncker, and Christopher D. Harvey. Representational drift: Emerging theories for continual learning and experimental future directions. *Current Opinion in Neurobiology*, 76:102609, October 2022.

[62] Leonhard Waschke, Niels A Kloosterman, Jonas Obleser, and Douglas D Garrett. Behavior needs neural variability. *Neuron*, 109(5):751–766, 2021.