An algorithmic hypothesis of differential neural stability in the brain

Kristopher T. Jensen@1

1Computational and Biological Learning Lab, Department of Engineering, University of Cambridge, Cambridge, U.K.

@ ktj21@cam.ac.uk

The animal brain has the ability to learn many new behaviors over the course of their lives. Despite this flexibility, most of the learned behaviors also remain highly stable after learning, even after long periods without continued practice. The stability of such remembered behaviors is seemingly at odds with the high degree of synaptic turnover observed in the brain (Holtmaat and Svoboda, 2009; Xu et al., 2009; Yang et al., 2009).

In agreement with a turnover of dendritic spines and other cellular components, many studies in neuroscience have found that task-associated neural representation drift over timescales of a few hours to days or weeks (Rokni et al., 2007; Carmena et al., 2005; Driscoll et al., 2017; Schoonover et al., 2020). Such unstable single-neuron representations have been hypothesized to facilitate stable behavior either by drifting in a functional 'null-space' (Gallego et al., 2020), or by limiting drift to directions that only require minor changes to a downstream neural decoder (Rule et al., 2020). However, under this hypothesis it remains an open question how neural circuits would identity such null- or decoder-limiting directions in which to drift.

On the other hand, a separate set of studies have suggested that neural representations of a given task tend to be stable after learning (Chestek et al., 2007; Flint et al., 2016; Dhawale et al., 2017). This is consistent with the observation of long-term stability in the zebra finch song circuit after initial learning in the juvenile bird (Katlowitz et al., 2018). However, in the context of zebra finch song learning, plasticity is limited to a single 'critical period' during development (Sizemore and Perkel, 2011) and the resulting circuit does not have to adapt to learning new behaviors. In contrast, the brain regions studied in mammals tend to undergo continual learning throughout life and thus must maintain the ability to learn and adapt in the face of these seemingly stable representations.

While the topic of stable neural representations and the apparent paradox with lifelong learning has long been a topic of interest and study in the neuroscience community, it has also recently begun to be addressed in the machine learning literature. In particular, while biological agents are capable of learning over a lifetime with little to no loss in performance on previously learned tasks, artificial agents tend to undergo 'catastrophic forgetting' whereby the performance on previous tasks deteriorates rapidly as new tasks are learned. This shortcoming of artificial agents has been addressed using methods ranging from 'replay' of examples from previous tasks (Van de Ven and Tolias, 2018; Pan et al., 2020; Li and Hoiem, 2017; Shin et al., 2017) to regularizing parameters important for previous tasks (Kirkpatrick et al., 2017; Ritter et al., 2018; Nguyen et al., 2017), and projecting parameter updates into subspaces that do not interfere with previous tasks (Zeng et al., 2019; Duncker et al., 2020).

In this short note, we attempt to relate experimental findings on neural stability to the machine learning literature and consider how qualitatively different approaches to addressing the continual learning problem can lead do different levels of representational stability at the single-neuron level. We discuss our results in light of experimental findings from different regions of the brain in terms of both the stability of the neural representations and experimental evidence for different mechanisms that might help overcome catastrophic forgetting.

Results

Task structure

To illustrate the implications of different continual learning algorithms on the stability of neural representations, we follow Kao et al. (2021) and consider a simple dataset which involves classifying sequential digits (de Jong, 2016). In contrast to most approaches to continual learning, we will work with a recurrent neural network models with dynamics given by

$$\mathbf{r}_t = \phi(\mathbf{A}\mathbf{r}_{t-1} + \mathbf{B}\mathbf{x}_t + \xi_t) = \phi(\mathbf{W}\mathbf{z}_t + \xi_t) \tag{1}$$

$$\mathbf{y}_t \sim p(\mathbf{y}_t | \mathbf{C}\mathbf{r}_t) \tag{2}$$

where we define $\mathbf{z}_t = (\mathbf{r}_{t-1}^\top, \mathbf{x}_t^\top)^\top$, $\mathbf{W} = (\mathbf{A}^\top, \mathbf{B}^\top)^\top$, and time is indexed by t. Here, $\mathbf{r} \in \mathbb{R}^{N_{rec} \times 1}$ are the network activations, $\mathbf{x} \in \mathbb{R}^{n_{in} \times 1}$ are the inputs, and $\mathbf{y} \in \mathbb{R}^{n_{out} \times 1}$ are the network outputs. We will define

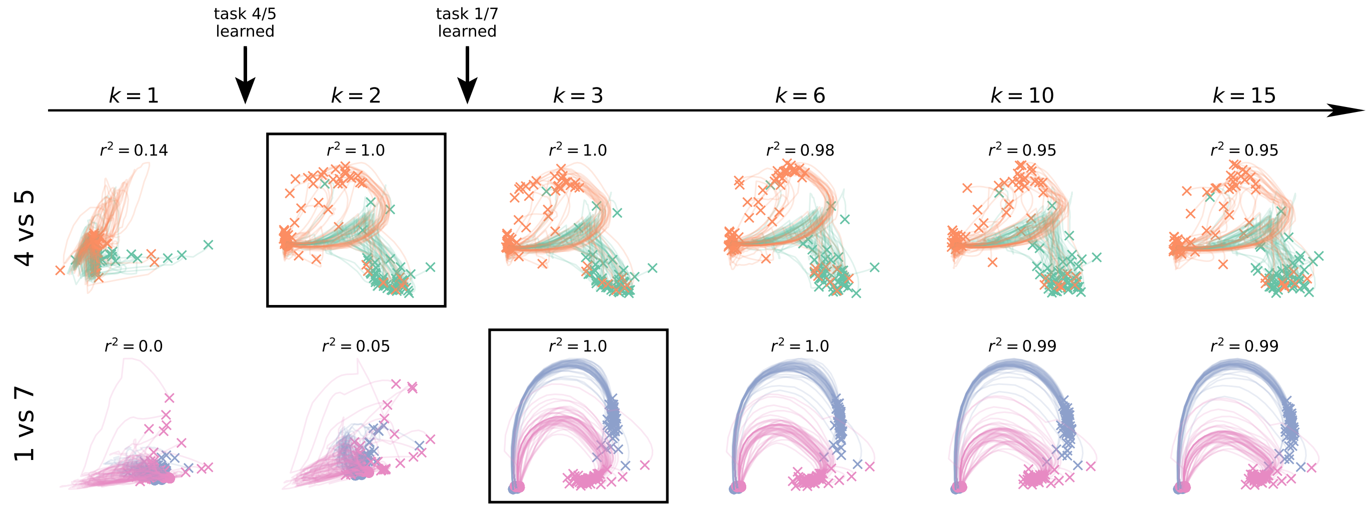

Figure 1: **Latent dynamics during SMNIST.** We considered two example tasks, 4 vs 5 (top) and 1 vs 7 (bottom). For each task, we simulated the response of a network trained by NCL to 100 digits drawn from that task distribution at different times during learning. We then fitted a factor analysis model for each example task to the response of the network right after the correponding task had been learned (squares; $k^* = 2$ and $k^* = 3$ respectively). We used this model to project the responses at different times during learning into a common latent space for each example task. For both example tasks, the network initially exhibited variable dynamics with no clear separation of inputs and subsequently acquired stable dynamics after learning to solve the task. The r^2 values above each plot indicate the similarity of neural population activity with that collected immediately after learning the corresponding task, quantified across all neurons (not just the 2D projection).

the expected loss on task k as

$$\ell_k(\theta) = \mathbb{E}\left[\sum_t \log p(\mathbf{y}_t^{(k)}|\mathbf{C}\mathbf{r}_t^{(k)})\right], \tag{3}$$

where $\mathbf{x}_t^{(k)}$ and $\mathbf{y}^{(k)}$ are drawn from the data distribution for task k and the expectation is computed at every iteration from a set of Monte Carlo samples. Continual learning in such recurrent neural networks has recently become a topic of interest in the machine learning community (Ehret et al., 2020; Duncker et al., 2020) and is of significant interest when trying to understand how catastrophic forggeting is mitigated in the brain; a network with a high degree of recurrency.

Continual learning algorithms

We will consider two broad classes of algorithms, namely those that regularize the *parameters* of the network and those that regularize the *input-output mapping* of the network. From the first class, we consider 'natural continual learning' (NCL; Kao et al., 2021) which regularizes the loss function on later tasks using the posterior over network parameters from previous tasks as a prior. This is combined with a form of gradient projection that encourages searching for local minima in the space of solutions to previous tasks to yield the following learning rule on task k:

$$\theta \leftarrow \theta - \gamma\left[\boldsymbol{\Lambda}_{k-1}^{-1}\nabla_\theta \ell_k(\theta) + (\theta - \mu_{k-1})\right] \tag{4}$$

where γ is a learning rate, θ are the network parameters, and $q_\phi(\theta) = \mathcal{N}(\theta; \mu_{k-1}, \boldsymbol{\Lambda}_{k-1}^{-1})$ is a Laplace approximation to the posterior over θ constructed from tasks 1 to k (see Kao et al., 2021 for details).

From the second class, we will consider a relatively naïve implementation of continual learning using replay. In this method, the learner estimates the task specific loss $\ell_k(\theta)$ as above. In addition, the learner gets to 'replay' a set of examples $\{\mathbf{x}^{(k')}, \mathbf{y}^{(k')}\}$ from previous tasks at every iteration to estimate the expected loss on earlier tasks

$$\ell_{<k}(\theta) = \frac{1}{k-1}\sum_{k'=1}^{k-1} \mathbb{E}\left[\sum_t \log p(\mathbf{y}_t^{(k')}|\mathbf{C}\mathbf{r}_t^{(k')})\right]. \tag{5}$$

The parameters are then updated as

$$\theta \leftarrow \theta - \gamma\left[\frac{1}{k}\nabla_\theta \ell_k(\theta) + \frac{k-1}{k}\nabla_\theta \ell_{<k}(\theta)\right]. \tag{6}$$

Note that while we explicitly replay examples drawn from the true data distribution for previous tasks, these examples can instead be drawn from a generative model that is learned in a continual fashion together with the discriminative model (Van de Ven and Tolias, 2018; van de Ven et al., 2020).

Task representations

We trained an RNN with 30 recurrent units on 15 sequential binary classification tasks from the extended SMNIST dataset used by Kao et al. (2021). We stored the parameters of the network after each task and simulated the response of the corresponding networks to a set

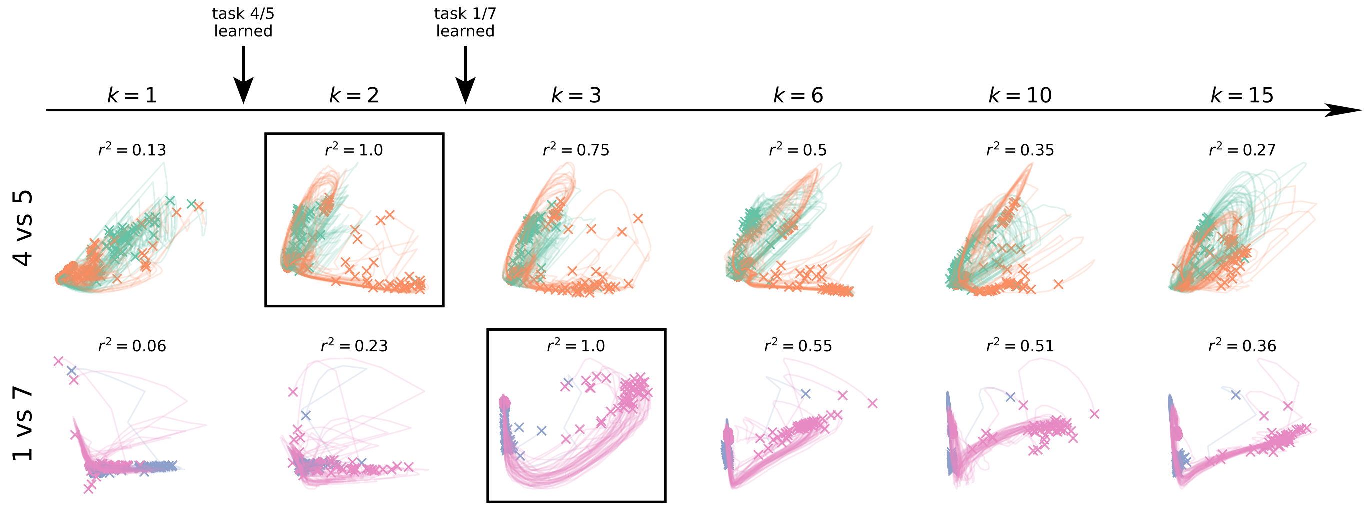

Figure 2: **Latent dynamics with replay.** As in Figure 1, now analyzing the dynamics of networks trained with replay.

of 100 digits drawn from the task distribution of classifying 4 vs 5 (task $k^* = 2$) or from classifying 1 vs 7 (task $k^* = 3$). This corresponds to returning to an old task ($k^* = 2$ or $k^* = 3$) at different stages of the sequential learning process ($k \in [1, 15]$). We projected the resulting dynamics for each task into a low-dimensional latent space by fitting a factor analysis model to $\{\mathbf{r}_t\}_1^T$ for the network specified by the parameters immediately after learning the corresponding task (i.e. $k = 2$ or $k = 3$). We then projected the dynamics after each task into the same latent space to visualize how they changed over time. This procedure reflects the approach used by Kao et al. (2021).

For networks trained by NCL, we found that the resulting latent trajecories changed with task learning but were subsequently stable (Figure 1). To quantify this at the single neuron level, we computed the correlation between the network dynamics $\{\mathbf{r}_t\}_1^T$ after each task $k \in [1, 15]$ and the dynamics right after the corresponding task k^* had been learned. Here we found a high degree of stability with an r^2 value of 0.95 for the 4 vs 5 task after learning an additional 13 tasks, and an r^2 value of 0.99 for the 1 vs 7 task after learning an additional 12 tasks. This is consistent with neuroscience studies suggesting a high degree of representational stability after learning (Dhawale et al., 2017; Chestek et al., 2007; Flint et al., 2016; Katlowitz et al., 2018). Furthermore, the learning rule employed by NCL regularizes changes in parameters from those used in previous tasks if they are 'important' for that task (c.f. Equation 4; see also Kirkpatrick et al., 2017). This is mechanistically similar to the observation of a stable subset of dendritic spines following task learning in previous experimental work (Yang et al., 2009; Fu et al., 2012) which lends some support to the implementation of mechanisms resembling parameter regularization for continual learning in biological systems.

When we trained the network using replay instead of NCL, the overall task performance was comparable. However, in contrast to NCL (and other weight regularization approaches such as EWC and KFAC), the networks trained with replay exhibited drifing representations of previously acquired tasks. When projecting the 30-dimensional dynamics into a low-dimensional space, we found that the network learned separable dynamics for each task, and that the two classes remained separable after further learning (Figure 2). However, the dynamics were no longer constrained to remain constant and instead drifted significantly during training with final r^2 values of 0.27 and 0.36 for the two tasks when compared to the representation immediately after task learning.

Local and global loss functions

We can understand the qualitative differences in task-associated dynamics between networks trained with NCL and replay by considering the locality of the loss function that is optimized. Consider an example loss function ℓ_1 for a hypothetical 'task 1' with two nearby local minima separated by a small energy barrier (Figure 3; left). When proceeding to train on task 2 after learning task 1, the replay-based approach to continual learning (Equation 5) provides an unbiased estimate of $\ell_1(\theta)$ although it may be very noisy if the number of replay events is small. This allows θ to move between adjacent local minima with task 1 performance, and the rate at which these transitions occur will increase with increasing noise level. If replay examples are instead drawn from a learned generative model, they are likely to provide a biased estimate of $\tilde{\ell}_1 \approx \ell_1$ but will still allow relatively uninhibited transitions between nearby minima of the approximate loss function.

Consider instead the Laplace approximation to $\ell_1(\theta)$ (Figure 3; right). In this case, the approximation $\tilde{\ell}_1 \approx \ell_1$ is inherently local and will always favour parameters

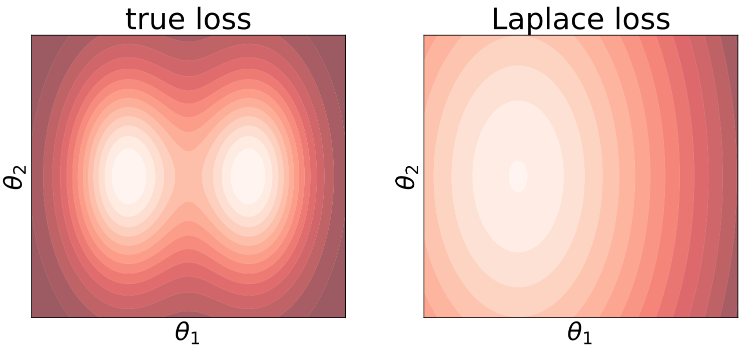

Figure 3: **Laplace approximation.** True loss for a hypothetical 'task 1' (left) and the approximate loss given by the Laplace approximation at one of two degenerate local minima.

θ that resemble the prior mean μ_1. This will break degeneracies between otherwise degenerate parameter sets and encourage a constancy of parameters that drives a constancy of dynamics. Indeed such a constancy of dynamics is directly built into a recent projection-based approach to continual learning by Duncker et al. (2020). A simple example of the degeneracy breaking introduced by the Laplace approximation is seen when permuting the identity of all recurrent units using some permutation operator $\hat{P}_u$. In the parameters θ are permuted accordingly, this will lead to a system with the exact same task performance. However, under the Laplace approximation, this new parameter set $\theta = \hat{P}_u(\mu)$ will have a much higher energy than the original parameters μ given the $||\theta - \mu||_2^2$ term of the Laplace loss function (see e.g. Huszár, 2017 for an explicit derivation of this loss). While we have discussed such degeneracy breaking for the Laplace approximation here, the same is true for any local approximation that restricts changes from the parameters used in a previous task or projects gradients based on the parameters or dynamics observed in previous task.

Discussion

In this short note we have highlighted how different continual learning algorithms can lead to qualitatively different properties of neural dynamics in recurrent networks over the course of learning. In particular, we have argued that continual learning algorithms based on local parameter regularization or projection matrices constructed from previous tasks will tend to preserve neural dynamics and lead to stable representations even as further tasks are learned. In contrast, algorithms that regularize the input-output mapping e.g. via exact or generative replay are susceptible to representational drift over the course of further learning.

It is interesting to speculate whether such algorithmic differences may explain some of the discrepancies in experimental reports of neural stability after task learning. Notably, several reports suggesting stable neural representations have centered around motor circuits (Dhawale et al., 2017; Chestek et al., 2007; Flint et al., 2016; Gallego et al., 2020) where previous work has also suggested that representations stabilize over learning (Ganguly and Carmena, 2009; Peters et al., 2014). This is consistent with our results for parameter regularized networks (Figure 1) which rely on restricting how much particular connections in the network can change depending on their importance for prior tasks. Importantly, experimental evidence for such regularization of specific connections has previously been reported in rodent motor cortex (Yang et al., 2009; Fu et al., 2012) suggesting a potential mechanistic explanation for the observed representational stability.

Conversely, drifting neural representations have been reported in several 'cognitive' regions including hippocampus (Ziv et al., 2013) and posterior parietal cortex (Driscoll et al., 2017; Rule et al., 2020). This is consistent with the important role of hippocampal replay in memory consolidation which is well-established in the neuroscience literature (Wilson and McNaughton, 1994; van de Ven et al., 2016; Carr et al., 2011). Additionally, replay-like events have been observed in neocortical regions including posterior parietal cortex (Qin et al., 1997). It is possible that such neural replay of past experiences provides a substrate for continual learning in these brain regions which obviates the need for stability at the level of single synapses and allows for continually drifting representations as in Figure 2.

Finally, we note that we have only considered 'replay' in the form of reproducing a set of inputs and target outputs $\{\mathbf{x}, \mathbf{y}\}$ from previous tasks with no constraints on the intermediate representations. However, it is also possible that replay events in the brain will constrain intermediate representations which would likely stabilize the post-learning dynamics more than the simple input-output replay considered in this note.

References

Carmena, J. M., Lebedev, M. A., Henriquez, C. S., and Nicolelis, M. A. (2005). Stable ensemble performance with single-neuron variability during reaching movements in primates. *Journal of Neuroscience*, 25(46):10712–10716.

Carr, M. F., Jadhav, S. P., and Frank, L. M. (2011). Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature neuroscience*, 14(2):147.

Chestek, C. A., Batista, A. P., Santhanam, G., Byron, M. Y., Afshar, A., Cunningham, J. P., Gilja, V., Ryu, S. I., Churchland, M. M., and Shenoy, K. V. (2007). Single-neuron stability during repeated reaching in macaque premotor cortex. *Journal of Neuroscience*, 27(40):10742–10750.

de Jong, E. D. (2016). Incremental sequence learning. *arXiv preprint arXiv:1611.03068*.

Dhawale, A. K., Poddar, R., Wolff, S. B., Normand, V. A., Kopelowitz, E., and Ölveczky, B. P. (2017). Automated long-term recording and analysis of neural activity in behaving animals. *Elife*, 6:e27702.

Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N., and Harvey, C. D. (2017). Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell*, 170(5):986–999.

Duncker, L., Driscoll, L., Shenoy, K. V., Sahani, M., and Sussillo, D. (2020). Organizing recurrent network dynamics by task-computation to enable continual learning. *Advances in Neural Information Processing Systems*, 33.

Ehret, B., Henning, C., Cervera, M. R., Meulemans, A., von Oswald, J., and Grewe, B. F. (2020). Continual learning in recurrent neural networks with hypernetworks. *arXiv preprint arXiv:2006.12109*.

Flint, R. D., Scheid, M. R., Wright, Z. A., Solla, S. A., and Slutzky, M. W. (2016). Long-term stability of motor cortical activity: implications for brain machine interfaces and optimal feedback control. *Journal of neuroscience*, 36(12):3623–3632.

Fu, M., Yu, X., Lu, J., and Zuo, Y. (2012). Repetitive motor learning induces coordinated formation of clustered dendritic spines in vivo. *Nature*, 483(7387):92–95.

Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A., and Miller, L. E. (2020). Long-term stability of cortical population dynamics underlying consistent behavior. *Nature neuroscience*, 23(2):260–270.

Ganguly, K. and Carmena, J. M. (2009). Emergence of a stable cortical map for neuroprosthetic control. *PLoS Biol*, 7(7):e1000153.

Holtmaat, A. and Svoboda, K. (2009). Experience-dependent structural synaptic plasticity in the mammalian brain. *Nature Reviews Neuroscience*, 10(9):647–658.

Huszár, F. (2017). On quadratic penalties in elastic weight consolidation. *arXiv preprint arXiv:1712.03847*.

Kao, T.-C., Jensen, K. T., Bernacchia, A., and Hennequin, G. (2021). Natural continual learning: success is a journey, not (just) a destination. *arXiv preprint arXiv:2106.08085*.

Katlowitz, K. A., Picardo, M. A., and Long, M. A. (2018). Stable sequential activity underlying the maintenance of a precisely executed skilled behavior. *Neuron*, 98(6):1133–1140.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Li, Z. and Hoiem, D. (2017). Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. (2017). Variational continual learning. *arXiv preprint arXiv:1710.10628*.

Pan, P., Swaroop, S., Immer, A., Eschenhagen, R., Turner, R. E., and Khan, M. E. (2020). Continual deep learning by functional regularisation of memorable past. *arXiv preprint arXiv:2004.14070*.

Peters, A. J., Chen, S. X., and Komiyama, T. (2014). Emergence of reproducible spatiotemporal activity during motor learning. *Nature*, 510(7504):263–267.

Qin, Y.-L., McNaughton, B. L., Skaggs, W. E., and Barnes, C. A. (1997). Memory reprocessing in corticocortical and hippocampocortical neuronal ensembles. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1360):1525–1533.

Ritter, H., Botev, A., and Barber, D. (2018). Online structured laplace approximations for overcoming catastrophic forgetting. *arXiv preprint arXiv:1805.07810*.

Rokni, U., Richardson, A. G., Bizzi, E., and Seung, H. S. (2007). Motor learning with unstable neural representations. *Neuron*, 54(4):653–666.

Rule, M. E., Loback, A. R., Raman, D. V., Driscoll, L. N., Harvey, C. D., and O'Leary, T. (2020). Stable task information from an unstable neural population. *Elife*, 9:e51121.

Schoonover, C. E., Ohashi, S. N., Axel, R., and Fink, A. J. (2020). Representational drift in primary olfactory cortex. *bioRxiv*.

Shin, H., Lee, J. K., Kim, J., and Kim, J. (2017). Continual learning with deep generative replay. *arXiv preprint arXiv:1705.08690*.

Sizemore, M. and Perkel, D. J. (2011). Premotor synaptic plasticity limited to the critical period for song learning. *Proceedings of the National Academy of Sciences*, 108(42):17492–17497.

van de Ven, G. M., Siegelmann, H. T., and Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):1–14.

Van de Ven, G. M. and Tolias, A. S. (2018). Generative replay with feedback connections as a general strategy for continual learning. *arXiv preprint arXiv:1809.10635*.

van de Ven, G. M., Trouche, S., McNamara, C. G., Allen, K., and Dupret, D. (2016). Hippocampal offline reactivation consolidates recently formed cell assembly patterns during sharp wave-ripples. *Neuron*, 92(5):968–974.

Wilson, M. A. and McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, 265(5172):676–679.

Xu, T., Yu, X., Perlik, A. J., Tobin, W. F., Zweig, J. A., Tennant, K., Jones, T., and Zuo, Y. (2009). Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature*, 462(7275):915–919.

Yang, G., Pan, F., and Gan, W.-B. (2009). Stably maintained dendritic spines are associated with lifelong memories. *Nature*, 462(7275):920–924.

Zeng, G., Chen, Y., Cui, B., and Yu, S. (2019). Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372.

Ziv, Y., Burns, L. D., Cocker, E. D., Hamel, E. O., Ghosh, K. K., Kitch, L. J., El Gamal, A., and Schnitzer, M. J. (2013). Long-term dynamics of ca1 hippocampal place codes. *Nature neuroscience*, 16(3):264.