NEURAL MANIFOLDS AND LEARNING REGIMES IN NEURAL-INTERFACE TASKS

ALEXANDRE PAYEUR[1,2,*], AMY L. ORSBORN[3,4,5] AND GUILLAUME LAJOIE[1,2,*]

[1]Department of Mathematics and Statistics, Université de Montréal, Montréal, QC, H3T 1J4, Canada
[2]Mila – Québec Artificial Intelligence Institute, Montréal, QC, H2S 3H1, Canada
[3]Department of Bioengineering, University of Washington, Seattle, WA 98195, USA
[4]Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98195, USA
[5]Washington National Primate Research Center, Seattle, WA 98195, USA
*Correspondence: alexandre.payeur@mila.quebec, g.lajoie@umontreal.ca

Abstract. Neural activity tends to reside on manifolds whose dimension is lower than the dimension of the whole neural state space. Experiments using brain-computer interfaces (BCIs) with microelectrode arrays implanted in the motor cortex of nonhuman primates have provided ways to test whether neural manifolds influence learning-related neural computations. Starting from a learned BCI-controlled motor task, these experiments explored the effect of changing the BCI decoder to implement perturbations that were either "aligned" or not with the pre-existing neural manifold. In a series of studies, researchers found that within-manifold perturbations (WMPs) evoked fast reassociations of existing neural patterns for rapid adaptation, while outside-manifold perturbations (OMPs) triggered a slower adaptation process that led to the emergence of new neural patterns. Together, these findings have been interpreted as suggesting that these different rates of adaptation might be associated with distinct learning mechanisms. Here, we investigated whether gradient-descent learning could alone explain these differences. Using an idealized model that captures the fixed-point dynamics of recurrent neural networks, we uncovered gradient-based learning dynamics consistent with experimental findings. Crucially, this experimental match arose only when the network was initialized in a lazier learning regime, a concept inherited from deep learning theory. A lazy learning regime—in contrast with a rich regime—implies small changes on synaptic strengths throughout learning. For OMPs, these small changes were less effective at increasing performance and could lead to unstable adaptation with a heightened sensitivity to learning rates. For WMPs, they helped reproduce the reassociation mechanism on short adaptation time scales, especially with large input variances. Since gradient descent has many biologically plausible variants, our findings establish lazy gradient-based learning as a plausible mechanism for adaptation under network-level constraints and unify several experimental results from the literature.

1. Introduction

The firing rates of an ensemble of N neurons naturally live in a N-dimensional neural space, with one dimension per neuron. The set of neural responses produced during the execution of a well-defined task define a manifold embedded in this neural space [1, 2]. The dimensionality of this manifold depends on task complexity, with simpler tasks leading to lower-dimensional manifolds [3, 4]. A low-dimensional manifold signifies the existence of a comparatively small set of dominant patterns of co-variability among neurons. In general, one could argue that structured neural activity is a reflection of the structure of the world [5], and this idea manifests itself in the existence of a continuous attractor encoding eye position [6], of object manifolds in deep convolutional networks modeling object recognition [7], of low-dimensional cognitive maps in hippocampal networks [8] and of manifolds for the control of movements [1]. In turn, structured activity is thought to bear on the efficient encoding of sensory information [9], the few-shot

learning of concepts [10] and generalization in timing tasks [11]. However, probing the multifaceted implications of neural plasticity on neural manifolds can be difficult without proper control over the behavioral degrees of freedom. Brain-computer interfaces (BCIs) provide such control by defining a direct, causal link between network activity and the behavioral output [12, 13]. Perturbing this link forces the brain to correct "misalignments" between neural representations and behaviors, providing information about the relationship between manifold structure and adaptation.

In a landmark study, Sadtler and co-workers [14] suggested that the difficulty with which new behaviors can be acquired depends on the underlying co-variability structure of neural activity. Learning new behaviors that require activity patterns outside a pre-existing manifold would be more difficult than learning behaviors that can repurpose this co-variability structure. They tested this hypothesis with a BCI that decoded neural activity from a pre-existing manifold to control a computer cursor and solve a center-out reaching task. The BCI allowed them to perturb the original mapping between neural activity and behavior such that the control space was either within the manifold (within-manifold perturbation [WMP]) or outside the manifold (outside-manifold perturbation [OMP]). Subjects could adapt rapidly, within a day, to WMPs (WM adaptation), but adaptation to OMPs (OM adaptation) was not possible during the same time frame. Subsequent work showed that OM adaptation requires days of training with a carefully crafted training protocol [15].

It has been argued that the learning mechanisms underlying WM and OM adaptation differ [15, 16]. Adaptation to WMPs is thought to rely on a fast reassociation of existing patterns from the manifold to new movement goals [17] whereas incremental adaptation to OMPs slowly generates new neural patterns [15]. These seemingly different forms of learning would underlie the distinct time scales of behavioral improvement between the two types of perturbation. These ideas further suggest that the learning mechanisms responsible for WM adaptation would not change the co-variability structure significantly, while mechanisms responsible for adaptation to OMPs would.

Recent theoretical works have begun to shed light on possible adaptation mechanisms underlying the WM/OM adaptation dichotomy [18, 19, 20]. Wärnberg and Kumar [18] described how a manifold can be implemented in functional spiking neural networks and suggested that adaptation to OMPs requires larger synaptic changes than adaptation to WMPs, and thus more learning. Feulner and Clopath [19] suggested that the disparity between WM and OM adaptation could be explained by a corruption of the feedback error signal driving the learning process. Here, we found instead that gradient-descent learning—which relies on uncorrupted feedback errors and possesses biologically-plausible counterparts (see, e.g., [21])—could explain this learning phenomenon without invoking extraneous learning signals or assuming different learning mechanisms for WM and OM adaptation. A recent, concurrent manuscript by Humphreys *et al.* [20] has argued along the same lines, using different methods. In our case, we designed a simplified noisy recurrent neural network and trained it on a BCI task emulating the experiments reported in Ref. [14]. A distinguishing feature of our study is that differences between the adaptation rates to WMPs and OMPs in agreement with experiments emerged only in one type of learning regime, as explained below.

Artificial neural network research has identified qualitatively different ways in which gradient-descent learning might find task solutions in weight space. The lazy (or kernel [22]) regime typically occurs when the initial weights are drawn from a distribution with high variance [23, 24]. In this case, internal neural representations do not need to change significantly to solve the task and, as a consequence, the network weights move little during learning. In contrast, the

rich (or feature-learning) regime is associated with low-variance initialization and the weights change significantly during learning to adapt representations (features) to task demands [25, 26]. These regimes not only affect learning dynamics [27], but also the generalization properties of the trained network [27, 23, 24, 28] as well as the structure of learned neural representations [29, 30, 31, 32].

Our model generated different WM and OM adaptation rates that reproduced experimental data. Crucially, this was only possible in a lazier regime whereas richer regimes led to key discrepancies with BCI experiments. In the lazier regime, different adaptation rates organically followed from interactions between the pre-existing neural co-variability and the posited learning rule. As in experiments, WM adaptation harnessed more of the pre-existing co-variability structure than OM adaptation, which tended to produce neural activity excursions away from it. Our further studies of adaptation to OMPs revealed that a slow learning-induced "stretching" of the activity covariance along the initial manifold underlay the creation of new neural patterns. Altogether, the constraints placed on learning by OMPs required a moderate amount of feature learning despite the high-variance, lazier-regime initialization, i.e., OM adaptation was effectively richer [33] than WM adaptation. We also discovered that OM adaptation was more unstable than WM adaptation, as assessed by following the maximum absolute eigenvalue of the recurrent weight matrix throughout adaptation. Finally, we found that our gradient-descent model could reproduce the reassociation mechanism [17] on short adaptation time scales, especially with large input variances. Together, our findings suggest that lazy-regime gradient-based learning might be a core principle behind adaptation to sensorimotor perturbations in a BCI context with important implications for the learning dynamics, in particular the magnitude of synaptic changes necessary to support adaptation.

2. Results

We designed our model to capture salient aspects of typical center-out reaching tasks with BCIs (Fig. 1A, top). In these, neural activity from motor cortex is recorded and fed to a decoder to drive a cursor on a 2D screen, from a center point to K equidistant targets on a circle. The experimental subject must learn to produce the correct neural activity patterns for a given decoder to accomplish the task and receive rewards. Here, we abstracted this process by training our network to produce outputs $\mathbf{u}$ matching K equally-spaced unit vectors representing the targets $\mathbf{d}_k = [\cos(2\pi k/K), \sin(2\pi k/K]^\mathsf{T}$, with $k = 0, \ldots, K-1$. Informally, we might interpret $\mathbf{u} = (u_x, u_y)$ as a cursor "trajectory" in the Cartesian plane and the $\mathbf{d}_k$'s as target "trajectories". The network sought to minimize the loss function (analogous to task error in BCI experiments)

$$L = \frac{1}{2K} \sum_{k=0}^{K-1} \mathbb{E}\left[\|\mathbf{u} - \mathbf{d}_k\|^2 \mid k\right], \tag{1}$$

which is the average over targets of the expected mean squared errors between the network's output and each target $\mathbf{d}_k$. The expectation $\mathbb{E}[\cdot|k]$ is conditioned on the target.

To show the adaptation phenomena in the most straightforward way, we studied a simple class of network models, which can be construed as the limit of a noisy recurrent neural network close to its fixed point (see Methods) [34]. The recurrent activity $\mathbf{v}$, a vector of size $N = 100$ to match loosely with the number of units recorded in [14], was given by

$$\mathbf{v} = (I - W)^{-1}(U\mathbf{x} + \boldsymbol{\xi}), \tag{2}$$

where U and W are the input and recurrent weight matrices, I is the identity matrix, $\boldsymbol{\xi}$ is a private (i.e., uncorrelated across units) zero-mean Gaussian noise applied to all recurrent units

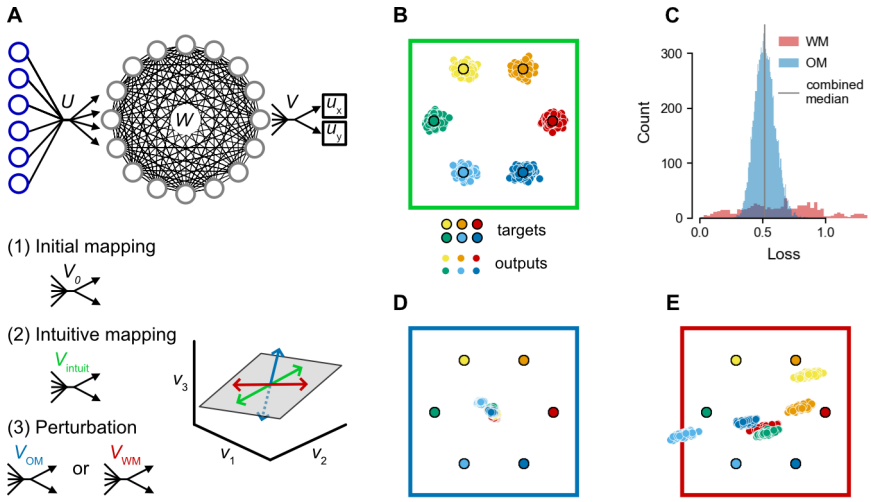

FIGURE 1. **Impact of the perturbations on the model's output before adaptation.** (A) Top: Schematic of the model. Bottom left: output matrix at different stages of an experiment. Bottom right (adapted from Ref. [14]): the different mappings depicted in a mock 3D neural space $\mathbf{v} = [v_1, v_2, v_3]^\mathsf{T}$. The intuitive mapping V_{intuit} (green double arrow, assuming a 1D output) lives in a manifold (grey plane) determined from the covariance of neural activity after training with $V = V_0$. The within-manifold perturbed mapping (V_{WM}, red double arrow) also belongs to that manifold while the outside-manifold perturbed mapping (V_{OM}, blue double arrow) does not. (B) Samples drawn from the model's output, $\mathbf{u} = (u_x, u_y)$, after fitting the decoder. (C) Distribution of candidate perturbations' losses. Vertical grey line indicates the median of the combined distributions. (D) Sampled outputs of the model before adaptation to the selected outside-manifold perturbation. (E) Same as D for the selected within-manifold perturbation.

and $\mathbf{x}$ is an input vector encoding information about the targets. The input $\mathbf{x}$ was a noiseless 1–of–K encoding—all elements were 0, except for a unique 1 at position k corresponding to target $\mathbf{d}_k$—with $K = 6$ targets, the private noise applied to each recurrent unit was low and only the recurrent weight matrix was plastic (see below). Elements of the input matrix U were fixed and randomly drawn from a uniform distribution between -1 and 1. Like other models [19], the output of the network was determined by a linear readout of the network activity

$$\mathbf{u} = V\mathbf{v}, \tag{3}$$

where V is an output matrix analogous to the BCI decoder used in experiments (e.g., when using a velocity Kalman filter, V would correspond to the Kalman gain matrix [14]).

This matrix adopted one of four possible matrix values (Fig. 1A, bottom):

(1) The initial mapping $V = V_0$ was a random Gaussian matrix (see Methods) meant to represent the readout during the experimental calibration blocks [14], prior to determining the intuitive mapping. The network was trained with $V = V_0$ until a good performance was achieved (see below).

(2) The intuitive mapping, denoted V_{intuit}, was defined as [19]

$$V_{\mathsf{intuit}} = DC.$$

The $M \times N$ matrix C projected network activity onto the leading $M = 6$ eigenvectors of the covariance matrix obtained after training was completed with $V = V_0$; the rows of C span the neural manifold (grey plane in Fig. 1A, bottom right). The $2 \times M$ matrix D was a linear readout connecting these projections to the output (see Methods). Right after fitting the intuitive mapping, outputs would be drawn from two-dimensional Gaussians mostly centered on the targets (Fig. 1B).

(3) The perturbed within-manifold (V_{WM}) and outside-manifold (V_{OM}) mappings were defined as follows. A WMP shuffled the rows of C, thus scrambling how the projections were fed to D, whereas an OMP shuffled the columns of C and altered how neural activity contributed to each pattern [14, 19]. Effectively, a WMP replaced V_{intuit} by a new output matrix

$$V_{\mathsf{WM}} = DP_{\mathsf{WM}}C, \tag{4}$$

where P_{WM} permutes the rows of C, while an OMP replaced V_{intuit} by

$$V_{\mathsf{OM}} = DCP_{\mathsf{OM}}, \tag{5}$$

where P_{OM} permutes the columns of C. As in experiments, after fitting the intuitive mapping, candidate WMPs and OMPs were tested by evaluating their effect on the immediate loss (Fig. 1C). As exemplified by the selected OMP (Fig. 1D), the OMPs produced outputs close to the center, resulting in a loss distribution with a small mean and a small variance; this was caused by the orientation of the perturbed mapping with respect to the intuitive mapping (Fig. 1A, bottom right). WMP losses were much more spread out, indicating that this type of perturbation rearranged the outputs more significantly, as shown for the selected WMP in Fig. 1E. Among all the candidate perturbations, the selected WMP and OMP were those whose effect on the immediate loss was closest to the median of the combined empirical distributions (Fig. 1C, grey line). This was done to make sure that adaptation to the WMP and OMP started from a similar loss.

To summarize, for each parameter initialization (random seed), the network was first trained using $V = V_0$. Then, the intuitive mapping was defined (which did not involved training the network). Finally, the selected WMP and OMP were applied in parallel and training was resumed starting from the network's configuration at the end of the initial training. At each stage—initial training and adaptation to the perturbations—only the recurrent weights W were learned by gradient descent on the loss function of Eq. 1, with updates given by

$$\Delta W = -\eta \nabla_W L, \tag{6}$$

where η is the learning rate hyperparameter which was fixed to 0.001 except stated otherwise. The rich vs lazy initialization of W will be described in the next section.

We reiterate that V was not learned through gradient descent but rather changed manually. In the following sections, V shall refer to the output matrix for either types of perturbation.

2.1. **Different adaptation rates to within- and outside-manifold perturbations in the lazy regime.** Our static model was valid only when the maximum absolute eigenvalue of W, $\max\{|\lambda(W)|\}$, stayed below 1 throughout learning because it represented the target-conditioned stable fixed points of a noisy and linear recurrent neural network (see Methods). We passively controlled this maximum eigenvalue by choosing Gaussian initializations of W (i.e., before the initial training with $V = V_0$) with standard deviation $N^{-\alpha}$ and varying α. We found empirically that 0.55 was the smallest α for which $\max\{|\lambda(W)|\} < 1$ during the adaptation process for the

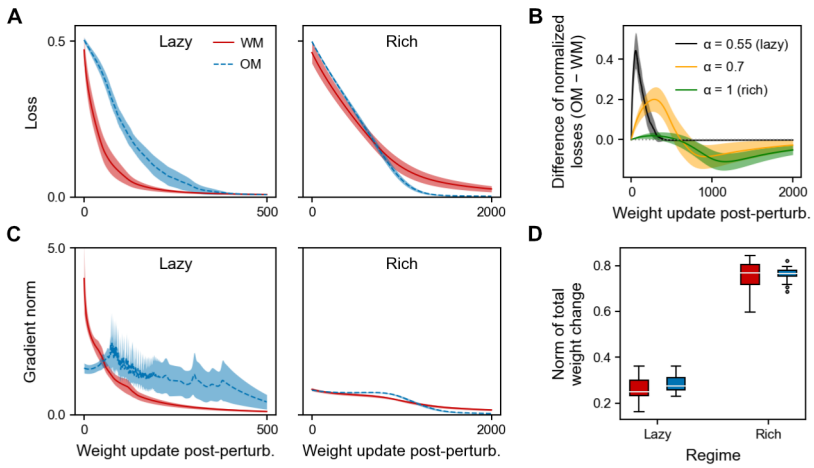

FIGURE 2. **Different adaptation rates to within- and outside-manifold perturbations in the lazy regime** (A) Loss during adaptation for the lazy (left; $\alpha = 0.55$) and rich (right; $\alpha = 1$) regimes. Note the different x axes. (B) Difference in normalized loss, where the normalized loss is L/L_0, with L_0 the loss at update 0. (C) Frobenius norm of the gradient, $\|\nabla_W L\|_F = \sqrt{\mathrm{tr}\,\{\nabla_W L \nabla_W L^\intercal\}}$, for the lazy (left) and rich (right) regimes. (D) Frobenius norm of the total weight change across adaptation. In panels A-C, we displayed mean $\pm 2 \times$ SEM, for $n = 20$ random seeds.

tested network initializations. In theory there was no upper bound on α, but in practice we restricted its range to $[0.55, 1]$. For brevity, we shall bend the machine learning terminology and refer to initializations with $\alpha = 0.55$ as high-variance or the lazy regime and initializations with $\alpha = 1$ as low-variance or the rich regime. Our "lazy" regime is actually the "laziest stable learning regime at the chosen learning rate with non-learnable input and output matrices".

The network was able to adapt completely to the two types of perturbation, both in the lazy and rich regimes (Figs. 2A). However, only for the lazy regime was OM adaptation slower than WM adaptation (Fig. 2A, left). In contrast, for the rich regime, the difference between adaptation to WMPs and OMPs nearly vanished during early adaptation and, later on, adaptation to OMPs was actually faster (Fig. 2A, right). Intermediate α's produced intermediate scenarios (Fig. 2B). Notwithstanding the difficulty of matching adaptation time scales measured in weight updates (model) to those measured in trials (experiments; see Discussion), these comparisons between WM and OM learning curves for different weight variances suggest that a lazy regime was necessary to produce slower OM adaptation.

Next, we needed to ascertain that the lazy and rich adaptation regimes were actually lazy and rich, irrespective of the perturbation type. Indeed, the variance of the initial weights—which determines the learning regime [24]—was specified before the initial training stage of the network. During this initial stage, the network learned to reach the targets, thus modifying the weight distribution for the next stage, i.e., the adaptation to the perturbations. High-variance initializations did give rise to larger gradients than those obtained with low-variance initializations, especially early in adaptation (Fig. 2C, compare left and right panels). Moreover, the magnitude of the total weight changes across learning were much lower for the lazy regime (Fig. 2D). Large gradients (meaning that small weight changes generated large loss changes) and small total

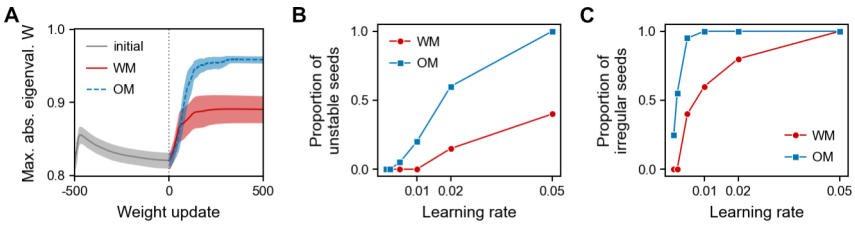

FIGURE 3. **Increased likelihood of learning-induced instability during adaptation to outside-manifold perturbations** (A) Maximum absolute eigenvalue of the recurrent weight matrix W during training with $\eta = 0.001$ (before and during adaptation; dotted vertical grey line indicates perturbation). We displayed mean $\pm$ 2 $\times$ SEM. (B) Proportion of unstable seeds (random initialization of network parameters), i.e. seeds for which $\max\{|\lambda(W)|\} > 1$ at the end of adaptation. (C) Proportion of irregular seeds (see text).

weight changes (meaning that the final solutions were close to the starting point in weight space) are hallmark of the lazy regime, *a posteriori* justifying our nomenclature for adaptation.

In accordance with the overall shape of the loss curves (Fig. 2A), gradient norms were much higher for adaptation to WMPs compared to OMPs early in adaptation in the lazy regime, thereafter becoming larger for OM adaptation, as well as more variable (Fig. 2C, left). Gradient norms in the rich regime were similar for both types of perturbation (Fig. 2C, right). Despite these different gradient norm profiles in the lazy regime, the total weight changes associated with WM and OM adaptation had similar magnitudes (Fig. 2D). Even though high-variance initializations indeed yielded lazy OM and WM adaptations, as assessed by the total weight change (Fig. 2D), the large difference in gradient norm early in adaptation suggests that WM adaptation is effectively lazier than OM adaptation, a notion to which we shall come back when analyzing the evolution of neural representations throughout adaptation (*cf.* Fig. 4).

The rest of the paper will focus on adaptation in the *lazy* regime.

2.2. **Increased likelihood of learning-induced instability during adaptation to outside-manifold perturbations.** In the previous section, we noticed that the gradient norms for OM adaptation were more variable than those for WM adaptation (Fig. 2C, left). This "noise" could be indicative of irregularities in the learning process following the onset of OMPs, despite the comparatively low learning rate ($\eta = 0.001$). While the maximum absolute eigenvalue of W, $\max\{|\lambda(W)|\}$, stayed below 1 during adaptation for both WM and OM adaptation, it rose to significantly higher values for the latter (Fig. 3A). For the same pre-perturbation W's, increasing the learning rate η increased both the propensity to generate $\max\{|\lambda(W)|\} > 1$ during adaptation (Fig. 3B) and the occurrence of irregular learning curves, defined as a loss that did not monotonically decrease during adaptation (Fig. 3C). Importantly, adaptation to OMPs was more sensitive to large learning rates, which was expected given the dynamics of $\max\{|\lambda(W)|\}$ during adaptation in Fig. 3A. The stress placed upon the network by OMPs thus tended to elicit more frequent excursions beyond the stability line.

2.3. **Slow covariance dynamics for outside-manifold adaptation.** To gain further insight into the different adaptation rates for WMPs and OMPs in the lazy regime, we rewrote the loss

(Eq. 1) as

$$L = \frac{1}{2} + \underbrace{\frac{1}{2}\mathsf{tr}\left\{V(\mathbb{V}[\mathbf{v}] + \bar{\mathbf{v}}\bar{\mathbf{v}}^{\mathsf{T}})V^{\mathsf{T}}\right\}}_{L_{\mathsf{corr}}} - \frac{1}{K}\sum_{k=0}^{K-1} \mathbf{d}_k^{\mathsf{T}} V \mathbb{E}[\mathbf{v}|k], \tag{7}$$

to decompose it into a contribution due to the output correlation, denoted L_{corr}, and a negative contribution penalizing any misalignment of the mean output with the target (Methods, Eq. 20). Importantly, L_{corr} does not refer to the targets, only to activity correlations projected onto the readout matrix V (see Discussion). Neglecting the contribution from the global mean activity $\bar{\mathbf{v}}$ (see Supp. Math Note 3 and Supp. Fig. 2), the correlation component of the loss suggests that the covariance $\mathbb{V}[\mathbf{v}]$ increasingly projects itself towards the perturbed decoder during adaptation (Fig. 4A). Indeed, $\mathbb{V}[\mathbf{v}]$ was given by (Methods, Eq. 16)

$$\mathbb{V}[\mathbf{v}] = \frac{1}{K}\sum_{k=0}^{K-1} (I-W)^{-1}\left[U(\boldsymbol{\mu}_k - \bar{\mathbf{x}})(\boldsymbol{\mu}_k - \bar{\mathbf{x}})^{\mathsf{T}}U^{\mathsf{T}} + \Sigma_\xi\right](I-W)^{-\mathsf{T}}, \tag{8}$$

where $\bar{\mathbf{x}} = (1/K)[1,\ldots,1]^{\mathsf{T}} \in \mathbb{R}^K$ is the global mean input, $\boldsymbol{\mu}_k$ is the kth 1–of–K encoding and Σ_ξ is the (diagonal) covariance of the private noise $\boldsymbol{\xi}$. Neglecting this private noise, we can write $L_{\mathsf{corr}} \approx (1/2K)\sum_k \|V(I-W)^{-1}U(\boldsymbol{\mu}_k - \bar{\mathbf{x}})\|^2$. For OMPs, L_{corr} is nearly zero right after the perturbation, meaning that the N-dimensional vectors $(I-W)^{-1}U(\boldsymbol{\mu}_k - \bar{\mathbf{x}})$ were mostly in the null space of V_{OM} (i.e., the set of vectors $\mathbf{y} \in \mathbb{R}^N$ such that $V_{\mathsf{OM}}\mathbf{y} = \mathbf{0}$). For WMPs, alignment of neural features started closer from the saturation value of 1/2, contributing to faster adaptation (Fig. 4A). This lesser strain on the activity covariance is intimately related to the lazier regime underlying WM adaptation (Fig. 2C, left and see Discussion). Also, given the behavior of L_{corr} during adaptation, we expect that further insight into the difference in adaptation under WMPs and OMPs would come from analyzing the covariance matrix $\mathbb{V}[\mathbf{v}]$ and, in particular, its interaction with the output matrix V.

Before adaptation, neural activity mostly resided in the low-dimensional manifold defined by the leading modes of the covariance matrix, and the intuitive mapping V_{intuit} directly exploited this aspect of neural activity. While relearning the center-out task following the perturbation, the fraction of the total variance within the manifold (Methods, Eq. 26) might change relative to what it was before the perturbation, denoting a learning-induced rearrangement of activity in neural space. The ratio of the fraction of variance explained by the intuitive manifold before and after learning has been called the normalized variance explained [17] or the manifold overlap [19] (Methods, Eq. 27). We found that the manifold overlap decreases from its starting value of 1 during adaptation for both types of perturbation, reaching significantly lower values for OMPs compare to WMPs (Fig. 4B). These results suggest that under gradient-based learning, a good portion of the adaptation strategy to OMPs relies on the pre-existing activity structure, but less so than WMPs.

Adaptation to OMPs was slower (Fig. 2A) than adaptation to WMPs, but the fraction of variance explained by the original manifold—represented by C—varied little for both types of perturbation (Fig. 4B). This begs the question: after an OMP, how did the network manage to project its covariance towards the output mapping V_{OM} while retaining most of its variance in the original manifold? Perhaps learning steers the variance towards the perturbed manifold CP_{OM} (see Eq. 5). We found that the projection onto CP_{OM} was not significantly altered relative to the projection onto C (Fig. 4C), but that its absolute value increased quite drastically during adaptation (Fig. 4D). These results suggest that adaptation to OMPs involved a large absolute (but small relative) reshaping of the covariance towards the subspace spanned by the orthogonal

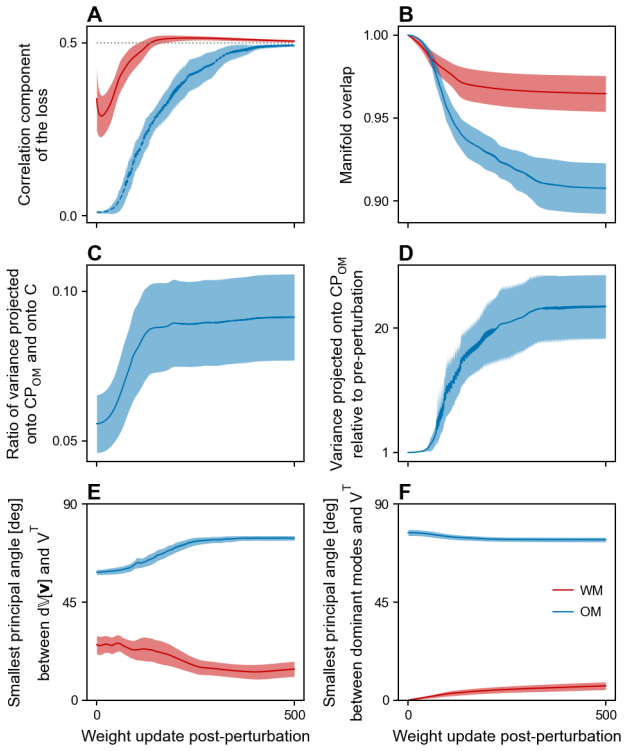

FIGURE 4. **Slow covariance dynamics for outside-manifold adaptation** (A) Correlation component of the loss. (B) Manifold overlap. (C) Ratio of projected variances for OMPs: $\mathsf{tr}\left\{C\mathbb{V}[\mathbf{v}]C^{\mathsf{T}}\right\}/\mathsf{tr}\left\{CP_{\mathsf{OM}}\mathbb{V}[\mathbf{v}]P_{\mathsf{OM}}^{\mathsf{T}}C^{\mathsf{T}}\right\}$ (Eq. 28). (D) Variance projected onto CP_{OM} relative to pre-perturbation. (E) Smallest principal angle between the subspaces spanned by the differential covariance $d\mathbb{V}[\mathbf{v}]$ and V^{T}. (F) Smallest principal angle between the subspaces spanned by the $M = 6$ leading modes of $\mathbb{V}[\mathbf{v}]$ and V^{T}. In all panels, we displayed mean $\pm 2 \times$ SEM, for $n = 20$ random seeds (random initialization of network parameters).

rows of CP_{OM} (see Methods), which is accomplished by stretching the variance inside the original manifold defined by C.

This picture was further confirmed by calculating the smallest principal angle [35] between the subspace spanned by the columns of $V_{\mathsf{OM}}^{\mathsf{T}}$ and the column space of the changes in covariance $\Delta\mathbb{V}[\mathbf{v}]$ throughout adaptation, which for small weight updates are well approximated by the differentials $d\mathbb{V}[\mathbf{v}]$. Here,

$$d\mathbb{V}[\mathbf{v}] = (I - W)^{-1} dW \mathbb{V}[\mathbf{v}] + \mathbb{V}[\mathbf{v}] dW^{\mathsf{T}} (I - W^{\mathsf{T}})^{-1}, \tag{9}$$

where $dW = \Delta W$ is the recurrent weight update generated by gradient descent (Eq. 6). (We note that the dependence of $\mathbb{V}[\mathbf{v}]$ (Eq. 8) on the output matrix V only occurs implicitly through the learning process by compounding differentials $d\mathbb{V}[\mathbf{v}]$ which depend on the weight updates.) The $d\mathbb{V}[\mathbf{v}]$'s are low-rank because the weight updates are low-rank (Methods, Eq. 24). During adaptation to OMPs, the subspace of the neural space spanned by the differential covariance was

$\sim 60^\circ - 75^\circ$ from the subspace spanned by $V_{\mathsf{OM}}^{\mathsf{T}}$ (Fig. 4E). For WMPs, these covariance updates were much more aligned with $V_{\mathsf{WM}}^{\mathsf{T}}$, with angles $\sim 15^\circ - 25^\circ$. The leading eigenmodes of $\mathbb{V}[\mathbf{v}]$ obeyed a similar principle, staying mostly away from $V_{\mathsf{OM}}^{\mathsf{T}}$ (Fig. 4F).

Figure 4 suggests the following crude illustration for the covariance dynamics during adaptation to OMPs. If $\mathbb{V}[\mathbf{v}]$ is pictured as a "vector" mostly orthogonal to the "vector" V_{OM} before adaptation (Fig. 4A, at 0 weight update post-perturbation), then each $d\mathbb{V}[\mathbf{v}]$ is a push applied on $\mathbb{V}[\mathbf{v}]$ with a small component parallel to V_{OM}, because $\angle(d\mathbb{V}[\mathbf{v}], V_{\mathsf{OM}}^{\mathsf{T}}) \in [60^\circ - 75^\circ]$ (Fig. 4E). Therefore, $\mathbb{V}[\mathbf{v}]$ slowly develops a projection onto CP_{OM} (Fig. 4D), but this projection is always small relative to the projection onto the original C (Figs. 4B and D). Adaptation to OMPs is thus mostly enacted by a slow and mildly anisotropic stretch of activity covariance in neural space, with a negligible rotation of its leading principal axes (Fig. 4F).

2.4. **Adaptation strategies to within-manifold perturbations.** Golub and co-workers [17] have suggested that adaptation to WMPs is mediated by a reassociation mechanism according to which a fixed repertoire of neural patterns is simply associated with different cursor movements during adaptation. Therefore, activity before and after adaptation should be similar. Alternative mechanisms like realignment and rescaling (see [17] and Methods) would involve more extensive changes to neural activity.

We have shown already that our neural network model adapts to WMPs with minimal changes to the fraction of variance projected onto the manifold generated by C (Fig. 4B). To reveal the adaptation mechanism underlying WM adaptation in our model, we followed Ref. [17] and computed the amount of co-variability projected onto the row space of the decoder D, before and after the WMP. If the strategy adopted by our model is reassociation, the projections onto both the original and perturbed decoders should be close to zero during adaptation. Since each random initialization of the network has a slightly different adaptation curve (i.e., loss vs weight update), we plotted these projection changes as a function of the percentage decrease in loss, to be able to evaluate the adaptation strategy at different learning stages (Fig. 5). Early during adaptation (50% percentage loss), our model was somewhat compatible with reassociation, albeit with a lot of variability across network initializations (Fig. 5A). Later during adaptation, the changes in projection increased drastically, away from a mere reassociation of patterns. Increasing the variance of the input received by the recurrent units (implemented in Fig. 5B by multiplying the 1–of-K encoding by 2) decreased both the medians and variances of the changes in projection and rendered the adaptation mechanism more compatible with reassociation. The converse—decreasing input variance—produced the opposite effect, with large and highly variable changes in projection throughout adaptation (Supp. Fig. 3A).

3. Discussion

Low-dimensional, task-induced neural manifolds might influence the propensity to learn subsequent tasks which harness activity from the neural population underlying these manifolds. BCI experiments have directly addressed this question, suggesting an inherent difficulty to adapt to perturbations which force a control space outside the pre-existing manifold [14, 17, 15]. In this paper, we used models to explore the hypothesis that a learning rule based on gradient descent could explain this phenomenon. We showed that the gap between WM and OM adaptation depended crucially on the learning regime (rich vs lazy), with only in the lazy(er) regime displaying marked differences in agreement with experiments (Fig. 2). Even though our network adapted to WMPs and OMPs by following the same learning rule, interactions between neural covariance and the perturbed mappings during learning generated contrasting adaptation signatures echoing (Fig. 4) and supplementing (Fig. 3) the experimental results. By providing precise

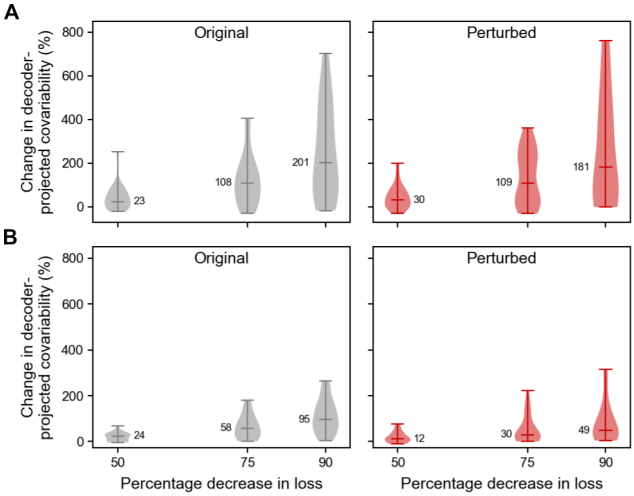

FIGURE 5. **Adaptation strategies to within-manifold perturbations.** (A) Change in decoder-projected co-variability onto the original (pre-perturbation; left) decoder and perturbed decoder (right). (B) Same as A, with increased input intensity. x-axis represents the percentage decrease with respect to the loss right after the WMP, corresponding to different stages of adaptation. Horizontal bars in the violin plots are medians over 20 random network initializations and their values are indicated next to them.

control over network parameters, our simple model suggests that what has been interpreted as different mechanisms [15, 16] may instead be described by a single learning computation within a particular network regime. Finally, we also studied adaptation strategies to WMPs, finding some support for a gradient-based reassociation mechanism early in training (Fig. 5).

3.1. **Alternatives to gradient descent.** Our work assumed that adaptation can be cast as an optimization problem [36] and that this optimization is carried out by gradient-descent learning. Gradient descent in neural networks relies on vector-based errors (*cf.* term $V\mathbb{E}[\mathbf{v}|k]-\mathbf{d}_k$ in Eq. 23), in contrast to other algorithms which require only a scalar reward signal, like REINFORCE [37]. However, because REINFORCE approximates gradient descent on average [37], we would expect our conclusions to hold at least qualitatively for that learning rule; the manuscript by Humphreys and co-workers [20] has shown that it is indeed the case for a feedforward neural network model. We would surmise that other learning rules which circumvent the biologically implausible characteristics of gradient descent [38] while still approximating its behavior should also agree qualitatively with our results. However, how well such rules align with gradient descent might depend on the learning regime itself [39] and thus might introduce additional details.

Further investigating alternative learning principles, we note that a key aspect of gradient descent is its scalar learning rate, in contrast to learning rules based on the recursive least-squares (RLS) algorithm. In the latter, the scalar learning rate is replaced by an estimate of an inverse correlation matrix [19, 40, 41]. This matrix learning rate implies that different directions in neural space have different effective learning rates or, equivalently, that the learning rates adapt as a function of the co-variability structure of neural activity (see Supp. Math Note 1 for more detail). Such adaptive learning rates might explain why adaptation to WMPs and OMPs are mostly identical when feedback error signals are uncorrupted (Supp. Fig. 1), justifying the need for corrupted error signals as was proposed in Ref. [19]. We note that a similar phenomenon

occurs in the rich learning regime where dominant features are learned faster than less dominant ones [42], contributing to the small difference between WM and OM adaptation.

An important difference between our results and those obtained with corrupted-feedback RLS-based adaptation is in the value of the manifold overlap (Fig. 4B). Feulner & Clopath obtained manifold overlaps in the range $\sim 25 - 75\%$ for WMPs and $\sim 2 - 60\%$ for OMPs, depending on the nature of the feedback error corruption [19]. In experiments, manifold overlaps for WM adaptation were measured to be at least 90% in 33/48 experiments and at least 74% in 47/48 experiments, or $93.4\% \pm 1.6\%$ (mean $\pm$ sem; see supplementary figure 3 in Ref. [17]; values extracted using WebPlotDigitizer [43]). Even though exact quantitative agreement would have been surprising given the simplicity of our model, we can conclude cautiously that it seems to agree at least qualitatively with experiments on WM adaptation, and more so than the model put forth in [19]. To our knowledge, manifold overlaps have not been computed yet for OMPs using experimental data. But we do know from Ref. [15], which focused on OMPs, that adaptation to OMPs combines within-manifold strategies and outside-manifold strategies, the first of which preserving the covariance structure and the second modifying it. In line with this finding, our model predicts that manifold overlap should be smaller for OM adaptation compared to WM adaptation, but not by much.

Although our research indicates that gradient descent in the lazy regime might play a significant role in generating the WM/OM adaptation dichotomy, we cannot exclude that other mechanisms are at play. In the context of BCI experiments where WMPs were applied, it has been suggested that changes in the internal state (e.g., motivation, engagement, arousal, etc.) affect how task performance improves post-perturbation [44]. However, the role played by such internal states in the differential adaptation to WMPs and OMPs has not been elucidated. It is possible that the perceived difficulty of OMPs—perhaps related to how quickly weight/synaptic changes are able to modulate motor outputs, i.e. the mechanism studied here—affects motivation and engagement in such a way that adaptation to OMPs becomes seemingly impossible on a single day, hence the need for a tedious multi-day learning curriculum to adapt to such perturbations [15].

3.2. **Rich versus lazy adaptation.** An important component of our findings was that conspicuous differences in the adaptation rates to WMPs and OMPs, as observed in experiments, appeared in the lazier learning regime but not the rich regime (Fig. 2A-B). Although the rich vs lazy dichotomy was clearly apparent in the norm of the total weight change across adaptation (Fig. 2D), for OM adaptation the gradient norms under the two regimes were not that dissimilar (Fig. 2C). In particular, the lazy OM gradient norm was significantly lower than the WM gradient norm early in training. Moreover, gradient-descent adaptation to WMPs yielded much easier and smaller feature alignment with the decoder (Fig. 4A) that harnessed significantly more of the pre-perturbation covariance structure (Fig. 4B). Feature alignment was possible because of the finite size of our network and because the weight variances were smaller than a pure lazy regime, in which no feature learning can occur [23]. In contrast, the network had to recruit activity patterns much beyond the unperturbed manifold to adapt to OMPs (Fig. 4B-D). It did so by stretching neural activity in neural space (Fig. 4C-F), increasing the likelihood of learning-induced instability (Fig. 3). Therefore, the slow adaptation to OMPs could perhaps be explained by a "conflict" between the high-variance, lazy, initialization and the requirement to learn new neural features, the latter being facilitated in richer regimes [24]. In sum, the learning regime could be said to be effectively lazier for WM adaptation compared to OM adaptation, a finding consistent with recent results arguing that an initial connectivity structure aligned with a task yields lazy-like learning dynamics [33].

We note that our definition for the lazy and rich regimes referred only to initialization of the recurrent weight matrix (W), ignoring the potential effect of the initialization of the input (U) and output matrices (V_0). Elements of the initial matrix V_0 scaled as $1/\sqrt{N}$ to facilitate comparison with the Feulner & Clopath model [19] (see Methods), but initializations scaling as $1/N$ are also possible, and could lead to different pre-perturbation operating points (*cf.* the recent preprint [45]).

As already mentioned, the lazier regime lead to small magnitudes of the total weight changes across adaptation, but there were no significant differences between OMP and WMP (Fig. 2D). This means that the fully adapted network after a WMP or an OMP ended up in (potentially) different local minima of the loss landscape at the same "distance" from the pre-adaptation weights, which they reached from different trajectories in weight space. This result differs from the study by Wärnberg and Kumar [18], which proposed that OM adaptation is slower because it requires larger changes in synaptic weights compared to WM adaptation. Feulner and Clopath [19], on the other hand, obtained similar weight changes for both types of adaptation, and suggested that the discrepancy between their (and our) findings and those of Wärnberg and Kumar originated from the way the perturbations were applied.

Assuming our conclusion about lazy adaptation to be valid, it is not completely clear why the motor system would pick the lazy regime. Several theoretical elements and confounding experimental factors might contribute to this. First, perhaps the BCI center-out task, with its small number of targets, two-dimensional movements, linear decoder, and no needs to generalize, is of low enough complexity that it warrants a local search in weight space. Second, the BCI experiments on which we based our study [14, 17, 15] involved animals that were highly trained on the task before the perturbations were applied, in contrast to naive animals performing a novel task. Third, recordings were made in the primary motor area [14], not in associative or prefrontal regions which likely play a role in learning and may showcase distinct learning dynamics. Fourth, perhaps metabolic constraints on plasticity are at play [46, 47, 48, 49, 50], since the lazy regime is associated with reduced weight changes [33] (Fig. 2D) that may be desirable for energy conservation. In support of our first three points, it has been suggested, in humans and macaques solving a context-dependent decision-making task, that the learned prefrontal/parietal cortical representations were compatible with a rich learning regime [31, 51] (see also [52]). In general, lazier learning solutions might offer generalization properties that could be advantageous in certain situations where reliance on multiple features is desirable [27]. Further investigation of this possibility using BCI task perturbations could yield important insight about learning strategies in the motor system.

3.3. **Time scales and performance metrics of adaptation.** It is difficult to compare the time scale of learning through exact gradient descent, which is measured in abstract weight updates, to that observed in experiments, which is measured in trials. An alternative to exact gradient descent (EGD) was to use samples drawn from the model instead of the exact distribution, which we shall call batch gradient descent (BGD) in a slight abuse of language. In BGD, the weights were updated after $K = 6$ trials, with a different target for each trial. BGD updates matched EGD updates quite precisely (Supp. Fig. 4). In Refs. [14, 17], which focused on a single day, adaptation was studied across $\sim 400-900$ trials (depending on the monkey and experiment). This translates to $\sim 67-150$ weight updates for EGD. The next hurdle to comparing theory and experiment quantitatively are the different performance metrics: acquisition times and success rates [14] or relative success rates [15] for experiments, and a loss function which accounts for the distance to targets for gradient descent. In experiments, on a single day, performance improved for WMPs but it did not recover its pre-perturbation level, especially in terms of the target

acquisition time (a full second of difference [14]); for OMPs, performance did not improve. In the model, the loss for OM adaptation did decrease over the $\sim [67-150]$ first weight updates (update 67: $69\% \pm 6\%$ [mean $\pm$ 2SEM]; update 150: $30\% \pm 6\%$), but how should one relate this decrease in loss to the absence thereof in the acquisition time-success rate coordinates? A better alternative would be to compare measures that, while correlated with performance, do not depend directly on movement goal. An example of such a measure would be L_{corr} (Eq. 7 and Fig. 4A), for instance by substituting for V the Kalman gain matrix of a velocity Kalman filter, when used as the BCI decoder (which is the case in all the experimental papers on which we based our work [14, 17, 15]). An easy conclusion from all these considerations is that future theoretical works should try to adopt the trial and reward structure of center-out BCI tasks to facilitate model-experiment comparisons.

3.4. **Stability of learning.** Our static recurrent network model became unstable whenever the maximum absolute eigenvalue of the recurrent weight matrix became greater than one during training. When this occurred, our model ceased to be the proper limit of the underlying dynamic recurrent network (see Methods) and the covariance could increase without bounds. At first sight, such instability could be construed as an artifact of our model, without neurobiological parallels. Indeed, homeostatic regulation of synaptic plasticity [53] and saturating firing rates could prevent such high-eigenvalue excursions of the weight matrix. Moreover, behaviors occur on finite time intervals so that the infinite-time limit never truly applies. But we would argue that these instabilities (Fig. 3B-C) or approaches to instability (Fig. 3A) can still serve to characterize the adaptation process. We interpreted the proportion of unstable and irregular seeds as representing how much stress both types of perturbation were applying on learning in the network. When the learning rate η increased in value, a clear difference appeared in these quantities between WM and OM adaptation. This suggests that the learning rate should be controlled more delicately for OM adaptation, which is perhaps what the incremental training protocol employed during experiments on multiple days was trying to accomplish indirectly [15]. However, this interpretation gives an important role to a model parameter which lacks a clear biological basis. Neuromodulatory centers innervating cortical circuits are routinely ascribed the role of modulating and controlling different aspects of cortical learning [54], but a more direct connection with η will require further research.

3.5. **Reassociation of patterns for within-manifold adaptation.** Our network model was not entirely compatible with the reassociation mechanism described by Golub and co-workers [17]. In that paper, the three experimental subjects (J, L and N) had changes in projections onto the original (intuitive) mapping of about -5%, 10% and -10% on average and changes in projections onto the perturbed mapping of about 5%, 5% and 10% on average, respectively. Standard deviations on these values were roughly 5%. We needed to upscale the input applied to the network to produced projection changes with a median of order 25% early in adaptation. Later in training, adaptation was in all cases more compatible with other mechanisms that change neural activity patterns more significantly (rescaling and realignment [see Methods]). In our model, only the recurrent connections were subject to plasticity while the input matrix was kept fixed. The first reason for keeping U unchanged was that the inputs were a simple 1–of–K encoding of target identity, not actual neural units. The second reason was to facilitate comparison with Ref. [19]—which used a fixed U—in order to shed light on the differences between adaptation mediated by a recursive least-squares and a gradient-descent learning rule. When we relaxed this constraint and made the input matrix the sole driver of adaptation (Supp. Fig. 3B) we obtained projection changes that were roughly comparable with those obtained with a fixed U and with the input multiplied by factor greater than that in Fig. 5B (Supp. Fig. 3C). Both these cases

were closer to the experimental results. An easy conclusion is that the input—its variance and plasticity—likely has a prominent role to play in the reassociation mechanism.

In summary, we showed that gradient-descent learning and the learning regime might be core components of adaptation to manifold-based perturbations in BCI tasks. In the lazier learning regime, neural co-variability interacted in a specific way with the sensorimotor mapping *via* the gradient-descent weight updates to generate different adaptation rates, providing a unified explanation of puzzling experimental phenomena.

4. Methods

4.1. **Network model.** The starting point of our model was the following discrete-time recurrent network dynamics

$$\begin{aligned} \mathbf{v}_{t+1} &= W\boldsymbol{\varphi}(\mathbf{v}_t) + U\mathbf{x} + \boldsymbol{\xi} \\ \mathbf{u}_t &= V\mathbf{v}_t, \end{aligned} \tag{10}$$

where $\mathbf{v}_t \in \mathbb{R}^N$ is the recurrent activity, $\mathbf{u}_t \in \mathbb{R}^2$ is the output at time t and $\boldsymbol{\varphi}(\cdot)$ is a nonlinear function applied element-wise (see below). Matrices W, U and V are the recurrent, input and output matrices, respectively. The noise $\boldsymbol{\xi}$ was Gaussian and uncorrelated across units, with diagonal covariance matrix $\Sigma_\xi = \sigma_\xi^2 I = 10^{-2} I$, abstracting a static neuronal noise source. The input $\mathbf{x}$ was modeled as a noiseless 1–of–K encoding—all elements were 0, except for a unique 1 at position k corresponding to target $\mathbf{d}_k$—with $K = 6$ targets. Time independence of both the input and the private noise is a drastic simplification, but allows for a simplified treatment of the fixed-point dynamics below.

To streamline modeling of the BCI phenomena under study, we assumed that the activation function $\boldsymbol{\varphi}(\cdot)$ was linear and we focused on the (noisy) fixed-point dynamics produced by Eq. 10. The solution to the linear dynamics with initial condition $\mathbf{v}_0$ is

$$\mathbf{v}_t = W^t \mathbf{v}_0 + (I - W)^{-1}(I - W^t)(U\mathbf{x} + \boldsymbol{\xi}),$$

assuming that $I - W$ is invertible. If the maximum absolute eigenvalue of W is less than 1, then $W^t \to 0$ as $t \to \infty$ and therefore

$$\lim_{t\to\infty} \mathbf{v}_t \triangleq \mathbf{v} = (I - W)^{-1}(U\mathbf{x} + \boldsymbol{\xi}), \tag{11}$$

while $\mathbf{u} = V\mathbf{v}$. Since every variable is Gaussian, the probability density of activity $\mathbf{v}$ is

$$p(\mathbf{v}) = \frac{1}{K}\sum_{k=0}^{K-1} p(\mathbf{v}|k) \tag{12}$$

where $p(\mathbf{v}|k)$ is Gaussian with mean $\mathbb{E}[\mathbf{v}|k]$ and covariance $\mathbb{V}[\mathbf{v}|k]$ given by

$$\mathbb{E}[\mathbf{v}|k] = (I - W)^{-1} U \boldsymbol{\mu}_k \tag{13}$$

$$\mathbb{V}[\mathbf{v}|k] = (I - W)^{-1} \Sigma_\xi (I - W^\mathsf{T})^{-1} \tag{14}$$

where $\boldsymbol{\mu}_k$ is the kth 1–of–K encoding.

The only adaptable matrix was W and its elements were initialized from a zero-mean Gaussian with variance $N^{-2\alpha}$, where α controlled the regime along the rich-lazy dimension. Elements of the input matrix U were randomly drawn from a uniform distribution between -1 and 1. The output matrix V will be described below.

4.2. **Task.** As stated in the main text, at any learning stage the objective was to minimize the loss function given by Eq. 1. We can expand this function to give (*cf.* Supp. Math Note 3)

$$L = \frac{1}{2K}\sum_k \mathsf{tr}\left\{V\mathbb{V}[\mathbf{v}|k]V^\mathsf{T}\right\} + \frac{1}{2K}\sum_k \|V\mathbb{E}[\mathbf{v}|k] - \mathbf{d}_k\|^2 \tag{15}$$

and make explicit the intuition that the network's goal was to reduce the average target-conditioned output variance and the distance between the target-conditioned mean output and $\mathbf{d}_k$. Yet another expression for the objective function makes use of the covariance matrix $\mathbb{V}[\mathbf{v}]$ which reads

$$\mathbb{V}[\mathbf{v}] = \frac{1}{K}\sum_k (I-W)^{-1}\left[U(\boldsymbol{\mu}_k - \bar{\mathbf{x}})(\boldsymbol{\mu}_k - \bar{\mathbf{x}})^\mathsf{T}U^\mathsf{T} + \Sigma_\xi\right](I-W)^{-\mathsf{T}}. \tag{16}$$

In this expression, $\bar{\mathbf{v}}$ is the global mean activity

$$\bar{\mathbf{v}} = (I-W)^{-1}U\bar{\mathbf{x}} \tag{17}$$

with $\bar{\mathbf{x}}$ the global mean input

$$\bar{\mathbf{x}} = \frac{1}{K}\sum_k \boldsymbol{\mu}_k = \frac{1}{K}[1,\ldots,1]^\mathsf{T}. \tag{18}$$

Multiplying Eq. 16 to the left by V, to the right by V^T and taking the trace yields (*cf.* Supp. Math Note 3)

$$\mathsf{tr}\left\{V\mathbb{V}[\mathbf{v}]V^\mathsf{T}\right\} = 2L - \mathsf{tr}\left\{V\bar{\mathbf{v}}\bar{\mathbf{v}}^\mathsf{T}V\right\} - 1 + \frac{2}{K}\sum_k \mathbf{d}_k^\mathsf{T}V\mathbb{E}[\mathbf{v}|k] \tag{19}$$

and therefore

$$L = \frac{1}{2} + \frac{1}{2}\mathsf{tr}\left\{V(\mathbb{V}[\mathbf{v}] + \bar{\mathbf{v}}\bar{\mathbf{v}}^\mathsf{T})V^\mathsf{T}\right\} - \frac{1}{K}\sum_k \mathbf{d}_k^\mathsf{T}V\mathbb{E}[\mathbf{v}|k]. \tag{20}$$

This expression means that the network will try to minimize the correlation matrix $\mathbb{V}[\mathbf{v}] + \bar{\mathbf{v}}\bar{\mathbf{v}}^\mathsf{T}$ (as "seen" from the output) and maximize the average overlap between the target and the target-conditioned expected output.

4.3. **Output matrix and BCI decoder.** We assumed that the dimensions of the output matrix V were such that $2 \ll N$, which is typically the case in BCI experiments ($\sim 85-91$ electrodes in [14], hence our selection of $N = 100$). The output matrix V adopted one of the four possible matrix values already mentioned in the main text; here, we simply complete the description of these matrices.

Initially, V was a random matrix V_0 with elements drawn from a standard normal distribution which were then multiplied by $0.57/\|V_0\|_F$ (see Ref. [19]; our prefactor of 0.57 differed from that article because of the different number of recurrent units). The symbol $\|\cdot\|_F$ denotes the Frobenius norm. Note that $\mathbb{E}[\|V_0\|_F] \sim \sqrt{N}$, and thus the elements of V_0 scale like $1/\sqrt{N}$. The network was then trained to solve the task with the gradient-descent algorithm described in section 4.5 below. At the end of this initial training, neural activity had variance $\mathbb{V}[\mathbf{v}]_0$ and global mean $\bar{\mathbf{v}}_0$, where the subscript $_0$ refers to using $V = V_0$.

The intuitive mapping was defined by $V_{\mathsf{intuit}} = DC$ [19]. Matrix D could be obtained by solving

$$D = \mathrm{argmin}_{D'}\mathbb{E}\left[\|(D'C - V_0)\mathbf{v}\|^2\right] \tag{21}$$

with solution (*cf.* Supp. Math Note 3)

$$D = V_0(\mathbb{V}[\mathbf{v}]_0 + \bar{\mathbf{v}}_0\bar{\mathbf{v}}_0^\mathsf{T})C^\mathsf{T}[C(\mathbb{V}[\mathbf{v}]_0 + \bar{\mathbf{v}}_0\bar{\mathbf{v}}_0^\mathsf{T})C^\mathsf{T}]^{-1}. \tag{22}$$

4.4. **Manifold's dimension and permutation selection.** The assigned dimensionality of the manifold, M, should be large enough to include the main neural modes [14], but small enough so that the WMP does not become as difficult as the OMP to relearn. The rule of thumb was that the chosen M should represent between 95% and 99% of the total variance. For the parameters used here, setting $M = 6$ allowed to represent most of the variance and kept the search for the WMPs computationally fast (see below). For different random seeds (different initializations of the network model), the number of principal components to reach 95% and 99% might differ (4.8 ± 0.1 for 95%; 7.2 ± 0.4 for 99% [mean $\pm$ SEM]), but M itself had the same value across seeds nevertheless.

There were $M!$ possible WMPs and $N!$ OMPs. Permutations were evaluated by computing their initial impact on the loss (with the pre-adaptation parameter values). Following Refs. [19, 14], for each type of perturbation we randomly generated a large number ($\mathsf{min}(M!, 10^4)$ for WMPs and 10^4 for OMPs) of candidate permutations, and then selected among these the permutation generating the loss closest to the median of the combined distributions, to make sure that adaptation to the WMP and OMP started from a similar loss.

4.5. **Learning.** For each training episode (with $V = V_0$ and then with either $V = V_{\mathsf{WM}}$ or $V = V_{\mathsf{OM}}$), gradient descent (Eq. 6) was used to minimize the loss (Eq. 1), with (Supp. Math Note 2)

$$\nabla_W L = (I - W^{\mathsf{T}})^{-1} V^{\mathsf{T}} \frac{1}{K} \sum_k \left[V \mathbb{V}[\mathbf{v}|k] + (V\mathbb{E}[\mathbf{v}|k] - \mathbf{d}_k)\mathbb{E}[\mathbf{v}|k]^{\mathsf{T}} \right]. \tag{23}$$

The sum over k combines a first term that relates to the output noise structure and a second term that involves the expected output error conditioned on the input component, $V\mathbb{E}[\mathbf{v}|k] - \mathbf{d}_k$. The sum is then pre-multiplied by $(I - W^{\mathsf{T}})^{-1} V^{\mathsf{T}}$ to recast its result into the neural space.

It is also possible to obtain an alternative expression for $\nabla_W L$ connected with Eq. 20, namely (*cf.* Supp. Math Note 2)

$$\nabla_W L = (I - W^{\mathsf{T}})^{-1} V^{\mathsf{T}} \left[V \left(\mathbb{V}[\mathbf{v}] + \bar{\mathbf{v}}\bar{\mathbf{v}}^{\mathsf{T}} \right) - \sum_k \pi_k \mathbf{d}_k \mathbb{E}[\mathbf{v}|k]^{\mathsf{T}} \right], \tag{24}$$

which emphasizes the impact of the covariance matrix $\mathbb{V}[\mathbf{v}]$ on the gradient.

During learning, weight updates produced changes to the covariance matrix, denoted $\Delta\mathbb{V}[\mathbf{v}]$. For small learning rates η, we can differentiate Eq. 16 and use $d(I-W)^{-1} = (I-W)^{-1} dW (I-W)^{-1}$ and $dW = \Delta W$ to approximate

$$\Delta\mathbb{V}[\mathbf{v}] \approx d\mathbb{V}[\mathbf{v}] = (I - W)^{-1} \Delta W \mathbb{V}[\mathbf{v}] + \mathbb{V}[\mathbf{v}] \Delta W^{\mathsf{T}} (I - W^{\mathsf{T}})^{-1}, \tag{25}$$

where $\mathbb{V}[\mathbf{x}]$ is the input covariance given by $\mathbb{V}[\mathbf{x}] = \frac{1}{K} \sum_k (\boldsymbol{\mu}_k - \bar{\mathbf{x}})(\boldsymbol{\mu}_k - \bar{\mathbf{x}})^{\mathsf{T}}$.

4.6. **Normalized explained variance / manifold overlap.** The matrix C projects the activity $\mathbf{v}$ onto the M principal orthogonal eigenvectors of the initial (i.e., before the perturbation) covariance matrix, denoted $\mathbb{V}[\mathbf{v}]_0$. Before adaptation, the fraction of explained variance by these projections was

$$f_0 = \frac{\mathbb{E}[\|C(\mathbf{v} - \bar{\mathbf{v}}_0)\|^2]}{\mathbb{E}[\|\mathbf{v} - \bar{\mathbf{v}}_0\|^2]} = \frac{\mathsf{tr}\left\{ C\mathbb{V}[\mathbf{v}]_0 C^{\mathsf{T}} \right\}}{\mathsf{tr}\left\{ \mathbb{V}[\mathbf{v}]_0 \right\}}.$$

During adaptation, the covariance might change and with it the fraction of variance explained by C, denoted f:

$$f = \frac{\mathsf{tr}\left\{ C\mathbb{V}[\mathbf{v}] C^{\mathsf{T}} \right\}}{\mathsf{tr}\left\{ \mathbb{V}[\mathbf{v}] \right\}}. \tag{26}$$

Intuitively, f will tend to increase if neural activity aligns more with C relative to the total variance, while f will decrease if misalignments become proportionally more prevalent. Therefore, the normalized variance explained [17] given by

$$\mathsf{NVE} = \frac{f}{f_0} \tag{27}$$

represented the relative proportion of the total variance that can be explained by the projection during adaptation. Feulner and Clopath [19] have referred to this quantity as the manifold overlap because this ratio represents the relative fraction of variance that lives in the row space of C. A manifold overlap of 100% thus means that the same proportion of variance (as compared to the pre-perturbation activity) lives in that space at that point during adaptation.

4.7. **Ratio of projected variances.** OMPs yielded a new C matrix, $C_{\mathsf{OM}} = CP_{\mathsf{OM}}$. Like the original C, the rows of C_{OM} are orthonormal, because $P_{\mathsf{OM}}P_{\mathsf{OM}}^{\mathsf{T}} = I \Rightarrow C_{\mathsf{OM}}C_{\mathsf{OM}}^{\mathsf{T}} = I$. Thus, as above, we can measure the intensity of the projected covariance within this subspace as $\mathsf{tr}\left\{C_{\mathsf{OM}}\mathbb{V}[\mathbf{v}]C_{\mathsf{OM}}^{\mathsf{T}}\right\}$. At each stage of the adaptation process, we can compute the ratio of projected variances

$$R = \frac{\mathsf{tr}\left\{C_{\mathsf{OM}}\mathbb{V}[\mathbf{v}]C_{\mathsf{OM}}^{\mathsf{T}}\right\}}{\mathsf{tr}\left\{C\mathbb{V}[\mathbf{v}]C^{\mathsf{T}}\right\}} \tag{28}$$

to quantify how adaptation to the OMP redistributes the covariance between the old (C) and new (C_{OM}) projections.

4.8. **Realignment, rescaling and reassociation.** Golub and co-workers [17] delineated three strategies whereby the motor cortex could adapt to WMPs. The initial decoder matrix D maps the projection of the activity $C\mathbf{v}$ onto "cursor positions". The WMP amounts to either leaving D unchanged and permuting the rows of C, or permuting the columns of D and leaving C unchanged. From the latter perspective, the WMP alters the two-dimensional row space $\subset \mathbb{R}^M$ of the decoder mapping. To define this row space, we can compute the (reduced) singular value decomposition (SVD) of the D matrix, $D = X_D S_D Y_D^{\mathsf{T}}$, where Y_D^{T} has two orthonormal rows spanning the row space of D. The amount of co-variability constrained to the unperturbed manifold span by C and projected along this row space can be computed as [17]

$$A(D) = \mathsf{tr}\left\{Y_D^{\mathsf{T}}C\mathbb{V}[\mathbf{v}]C^{\mathsf{T}}Y_D\right\}.$$

This quantity may vary throughout learning for both the original and perturbed mappings, which we denote by $A(D)$ and $A(DP_{\mathsf{WM}})$. *Realignment* means that the magnitude of the projections along the original row space decreases while those along the perturbed mapping increases, suggesting the activity changes to aligns itself to the new mapping. *Rescaling* consists in an increase of $A(D)$ during adaptation while $A(DP_{\mathsf{WM}})$ mostly does not change, suggesting that the network "stretches" its activity along the original mapping to compensate for the perturbation. Golub and co-workers suggested that it is a third strategy, *reassociation*, that best explained the experimental data. The name implies that the network relies on a fixed repertoire of neural patterns, which are repurposed following the perturbation. This reassociation means that both $A(D)$ and $A(DP_{\mathsf{WM}})$ barely change during learning. Following Ref. [17], we computed the relative change in decoder-projected co-variability (in %) as

$$\delta_{\mathsf{rel.}}A(\mathcal{D}) = 100\frac{A(\mathcal{D}) - A_0(\mathcal{D})}{A_0(\mathcal{D})} \tag{29}$$

where $\mathcal{D}$ is either D for the covariance projected onto the original decoder or DP_{WM} for the co-variance projected onto the perturbed decoder. The change is measured relative to the projected covariance right before the perturbation was applied, denoted $A_0(\mathcal{D})$. For realignment, rescaling

and reassociation, the pairs $(\delta_{\mathsf{rel.}} A(D), \delta_{\mathsf{rel.}} A(DP_{\mathsf{WM}}))$ should be (small negative, large positive), (large positive, ≈ 0) and (≈ 0, ≈ 0), respectively.

Acknowledgments

We wish to thank Yuhan Helena Liu for helpful comments on the manuscript. The authors' research was supported in part by an IVADO postdoctoral fellowship, Canada First Research Excellence Fund/Apogée (AP), the Simons Foundation (GL and ALO), a Google Faculty Award (ALO and GL) and an NSERC Discovery Grant (GL). GL further acknowledges support from the Canada CIFAR AI Chair Program and the Canada Research Chair in Neural Computations and Interfacing (CIHR, tier 2). The content of the present paper is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Author contributions

AP developed the model with inputs from GL, performed simulations and drafted the article. GL supervised the project. AP, ALO and GL revised the article.

Competing interests

ALO is a scientific advisor for Meta Reality Labs. GL is a scientific advisor for BIOS Health. AP declares no competing interests.

References

[1] J. A. Gallego, M. G. Perich, L. E. Miller, and S. A. Solla, "Neural Manifolds for the Control of Movement," *Neuron*, vol. 94, pp. 978–984, June 2017.

[2] M. Jazayeri and A. Afraz, "Navigating the Neural Space in Search of the Neural Code," *Neuron*, vol. 93, pp. 1003–1014, Mar. 2017.

[3] P. Gao and S. Ganguli, "On simplicity and complexity in the brave new world of large-scale neuroscience," *Current Opinion in Neurobiology*, vol. 32, pp. 148–155, June 2015.

[4] P. Gao, E. Trautmann, B. Yu, G. Santhanam, S. Ryu, K. Shenoy, and S. Ganguli, "A theory of multineuronal dimensionality, dynamics and measurement," tech. rep., Nov. 2017. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

[5] H. S. Seung and D. D. Lee, "The Manifold Ways of Perception," *Science*, vol. 290, pp. 2268–2269, Dec. 2000. Publisher: American Association for the Advancement of Science.

[6] H. S. Seung, "How the brain keeps the eyes still," *Proceedings of the National Academy of Sciences*, vol. 93, no. 23, pp. 13339–13344, 1996.

[7] U. Cohen, S. Chung, D. D. Lee, and H. Sompolinsky, "Separability and geometry of object manifolds in deep neural networks," *Nat Commun*, vol. 11, p. 746, Feb. 2020. Number: 1 Publisher: Nature Publishing Group.

[8] E. H. Nieh, M. Schottdorf, N. W. Freeman, R. J. Low, S. Lewallen, S. A. Koay, L. Pinto, J. L. Gauthier, C. D. Brody, and D. W. Tank, "Geometry of abstract learned knowledge in the hippocampus," *Nature*, vol. 595, pp. 80–84, July 2021. Number: 7865 Publisher: Nature Publishing Group.

[9] C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, and K. D. Harris, "High-dimensional geometry of population responses in visual cortex," *Nature*, vol. 571, pp. 361–365, July 2019. Number: 7765 Publisher: Nature Publishing Group.

[10] B. Sorscher, S. Ganguli, and H. Sompolinsky, "Neural representational geometry underlies few-shot concept learning," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 119, p. e2200800119, Oct. 2022.

[11] M. Beiran, N. Meirhaeghe, H. Sohn, M. Jazayeri, and S. Ostojic, "Parametric control of flexible timing through low-dimensional neural manifolds," *Neuron*, Jan. 2023.

[12] M. D. Golub, S. M. Chase, A. P. Batista, and B. M. Yu, "Brain–computer interfaces for dissecting cognitive processes underlying sensorimotor control," *Current Opinion in Neurobiology*, vol. 37, pp. 53–58, Apr. 2016.

[13] A. L. Orsborn and B. Pesaran, "Parsing learning in networks using brain–machine interfaces," *Current Opinion in Neurobiology*, vol. 46, pp. 76–83, Oct. 2017.

[14] P. T. Sadtler, K. M. Quick, M. D. Golub, S. M. Chase, S. I. Ryu, E. C. Tyler-Kabara, B. M. Yu, and A. P. Batista, "Neural constraints on learning," *Nature*, vol. 512, pp. 423–426, Aug. 2014.

[15] E. R. Oby, M. D. Golub, J. A. Hennig, A. D. Degenhart, E. C. Tyler-Kabara, B. M. Yu, S. M. Chase, and A. P. Batista, "New neural activity patterns emerge with long-term learning," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 116, pp. 15210–15215, July 2019.

[16] J. A. Hennig, E. R. Oby, D. M. Losey, A. P. Batista, B. M. Yu, and S. M. Chase, "How learning unfolds in the brain: toward an optimization view," *Neuron*, vol. 109, pp. 3720–3735, Dec. 2021.

[17] M. D. Golub, P. T. Sadtler, E. R. Oby, K. M. Quick, S. I. Ryu, E. C. Tyler-Kabara, A. P. Batista, S. M. Chase, and B. M. Yu, "Learning by neural reassociation," *Nature Neuroscience*, vol. 21, pp. 607–616, Apr. 2018.

[18] E. Wärnberg and A. Kumar, "Perturbing low dimensional activity manifolds in spiking neuronal networks," *PLOS Computational Biology*, vol. 15, p. e1007074, May 2019. Publisher: Public Library of Science.

[19] B. Feulner and C. Clopath, "Neural manifold under plasticity in a goal driven learning behaviour," *PLOS Computational Biology*, vol. 17, p. e1008621, Feb. 2021. Publisher: Public Library of Science.

[20] P. C. Humphreys, K. Daie, K. Svoboda, M. Botvinick, and T. P. Lillicrap, "BCI learning phenomena can be explained by gradient-based optimization," Dec. 2022. Pages: 2022.12.08.519453 Section: New Results.

[21] J. C. R. Whittington and R. Bogacz, "Theories of Error Back-Propagation in the Brain," *Trends in Cognitive Sciences*, vol. 23, pp. 235–250, Mar. 2019.

[22] A. Jacot, F. Gabriel, and C. Hongler, "Neural Tangent Kernel: Convergence and Generalization in Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018.

[23] L. Chizat, E. Oyallon, and F. Bach, "On Lazy Training in Differentiable Programming," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[24] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro, "Kernel and Rich Regimes in Overparametrized Models," in *Proceedings of Thirty Third Conference on Learning Theory*, pp. 3635–3673, PMLR, July 2020. ISSN: 2640-3498.

[25] G. Yang and E. J. Hu, "Tensor Programs IV: Feature Learning in Infinite-Width Neural Networks," in *Proceedings of the 38th International Conference on Machine Learning*, pp. 11727–11737, PMLR, July 2021. ISSN: 2640-3498.

[26] B. Bordelon and C. Pehlevan, "Self-Consistent Dynamical Field Theory of Kernel Evolution in Wide Neural Networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32240–32256, Dec. 2022.

[27] T. George, G. Lajoie, and A. Baratin, "Lazy vs hasty: linearization in deep networks impacts learning schedule based on example difficulty," *Trans. Mach. Learn. Res.*, vol. 2022, 2022.

[28] M. Geiger, S. Spigler, A. Jacot, and M. Wyart, "Disentangling feature and lazy training in deep neural networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2020, p. 113301, Nov. 2020. Publisher: IOP Publishing and SISSA.

[29] A. M. Saxe, J. L. McClelland, and S. Ganguli, "A mathematical theory of semantic development in deep neural networks," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 11537–11546, June 2019. Publisher: Proceedings of the National Academy of Sciences.

[30] J. Paccolat, L. Petrini, M. Geiger, K. Tyloo, and M. Wyart, "Geometric compression of invariant manifolds in neural networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, p. 044001, Apr. 2021. Publisher: IOP Publishing and SISSA.

[31] T. Flesch, K. Juechems, T. Dumbalska, A. Saxe, and C. Summerfield, "Orthogonal representations for robust context-dependent task performance in brains and neural networks," *Neuron*, vol. 110, pp. 1258–1270.e11, Apr. 2022.

[32] M. Farrell, S. Recanatesi, and E. Shea-Brown, "From lazy to rich to exclusive task representations in neural networks and neural codes," *Current Opinion in Neurobiology*, vol. 83, p. 102780, Dec. 2023.

[33] Y. H. Liu, A. Baratin, J. Cornford, S. Mihalas, E. Shea-Brown, and G. Lajoie, "How connectivity structure shapes rich and lazy learning in neural circuits," *ArXiv*, p. arXiv:2310.08513v1, Oct. 2023.

[34] G. Hennequin, Y. Ahmadian, D. B. Rubin, M. Lengyel, and K. D. Miller, "The Dynamical Regime of Sensory Cortex: Stable Dynamics around a Single Stimulus-Tuned Attractor Account for Patterns of Noise Variability," *Neuron*, vol. 98, pp. 846–860.e5, May 2018.

[35] Å. Björck and G. H. Golub, "Numerical methods for computing angles between linear subspaces," *Math. Comp.*, vol. 27, no. 123, pp. 579–594, 1973.

[36] B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, C. J. Gillon, D. Hafner, A. Kepecs, N. Kriegeskorte, P. Latham, G. W. Lindsay, K. D. Miller, R. Naud, C. C. Pack, P. Poirazi, P. Roelfsema, J. Sacramento, A. Saxe, B. Scellier, A. C. Schapiro, W. Senn, G. Wayne, D. Yamins, F. Zenke, J. Zylberberg, D. Therien, and K. P. Kording, "A deep learning framework for neuroscience," *Nature Neuroscience*, vol. 22, pp. 1761–1770, Nov. 2019. Number: 11 Publisher: Nature Publishing Group.

[37] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach Learn*, vol. 8, pp. 229–256, May 1992.

[38] F. Crick, "The recent excitement about neural networks," *Nature*, vol. 337, pp. 129–132, Jan. 1989. Number: 6203 Publisher: Nature Publishing Group.

[39] B. Bordelon and C. Pehlevan, "The Influence of Learning Rule on Representation Dynamics in Wide Neural Networks," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, OpenReview.net, 2023.

[40] D. Sussillo and L. Abbott, "Generating Coherent Patterns of Activity from Chaotic Neural Networks," *Neuron*, vol. 63, pp. 544–557, Aug. 2009.

[41] R. Laje and D. V. Buonomano, "Robust timing and motor patterns by taming chaos in recurrent neural networks," *Nature Neuroscience*, vol. 16, pp. 925–933, July 2013. Number: 7 Publisher: Nature Publishing Group.

[42] A. Baratin, T. George, C. Laurent, R. D. Hjelm, G. Lajoie, P. Vincent, and S. Lacoste-Julien, "Implicit Regularization via Neural Feature Alignment," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 2269–2277, PMLR, Mar. 2021. ISSN: 2640-3498.

[43] A. Rohatgi, "Webplotdigitizer: Version 4.6," 2022.

[44] J. A. Hennig, E. R. Oby, M. D. Golub, L. A. Bahureksa, P. T. Sadtler, K. M. Quick, S. I. Ryu, E. C. Tyler-Kabara, A. P. Batista, S. M. Chase, and B. M. Yu, "Learning is shaped by abrupt changes in neural engagement," *Nature Neuroscience*, vol. 24, pp. 727–736, May 2021. Number: 5 Publisher: Nature Publishing Group.

[45] F. Schuessler, F. Mastrogiuseppe, S. Ostojic, and O. Barak, "Aligned and oblique dynamics in recurrent neural networks," Sept. 2023. arXiv:2307.07654 [q-bio].

[46] F. Mery and T. J. Kawecki, "A Cost of Long-Term Memory in Drosophila," *Science*, vol. 308, pp. 1148–1148, May 2005. Publisher: American Association for the Advancement of Science.

[47] P.-Y. Placcais and T. Preat, "To Favor Survival Under Food Shortage, the Brain Disables Costly Memory," *Science*, vol. 339, pp. 440–442, Jan. 2013. Publisher: American Association for the Advancement of Science.

[48] P.-Y. Placcais, É. de Tredern, L. Scheunemann, S. Trannoy, V. Goguel, K.-A. Han, G. Isabel, and T. Preat, "Upregulated energy metabolism in the Drosophila mushroom body is the trigger for long-term memory," *Nature Communications*, vol. 8, p. 15510, June 2017. Number: 1 Publisher: Nature Publishing Group.

[49] H. L. Li and M. C. van Rossum, "Energy efficient synaptic plasticity," *eLife*, vol. 9, p. e50804, Feb. 2020. Publisher: eLife Sciences Publications, Ltd.

[50] J. Karbowski, "Metabolic constraints on synaptic learning and memory," *Journal of Neurophysiology*, vol. 122, pp. 1473–1490, Oct. 2019. Publisher: American Physiological Society.

[51] T. Flesch, A. Saxe, and C. Summerfield, "Continual task learning in natural and artificial agents," *Trends in Neurosciences*, vol. 46, pp. 199–210, Mar. 2023.

[52] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi, "The importance of mixed selectivity in complex cognitive tasks," *Nature*, vol. 497, pp. 585–590, May 2013. Number: 7451 Publisher: Nature Publishing Group.

[53] H.-K. Lee and A. Kirkwood, "Mechanisms of Homeostatic Synaptic Plasticity in vivo," *Frontiers in Cellular Neuroscience*, vol. 13, 2019.

[54] K. Doya, "Metalearning and neuromodulation," *Neural Networks*, vol. 15, pp. 495–506, June 2002.