Stochastic Gradient Descent-Induced Drift of Representation in a Two-Layer Neural Network

Farhad Pashakhanloo [1]   Alexei Koulakov [1]

Abstract

Representational drift refers to over-time changes in neural activation accompanied by a stable task performance. Despite being observed in the brain and in artificial networks, the mechanisms of drift and its implications are not fully understood. Motivated by recent experimental findings of stimulus-dependent drift in the piriform cortex, we use theory and simulations to study this phenomenon in a two-layer linear feedforward network. Specifically, in a continual online learning scenario, we study the drift induced by the noise inherent in the Stochastic Gradient Descent (SGD). By decomposing the learning dynamics into the normal and tangent spaces of the minimum-loss manifold, we show the former corresponds to a finite variance fluctuation, while the latter could be considered as an effective diffusion process on the manifold. We analytically compute the fluctuation and the diffusion coefficients for the stimuli representations in the hidden layer as functions of network parameters and input distribution. Further, consistent with experiments, we show that the drift rate is slower for a more frequently presented stimulus. Overall, our analysis yields a theoretical framework for better understanding of the drift phenomenon in biological and artificial neural networks.

1. Introduction

Representational drift has been observed across different parts of the nervous system, such as in the hippocampus (Ziv et al., 2013), sensorimotor (Rule et al., 2019), and visual systems (Deitch et al., 2021; Marks & Goard, 2021). A recent study in the piriform cortex (Schoonover et al., 2021), a

---

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, NY, USA. Correspondence to: Farhad Pashakhanloo <pashakh@cshl.edu>.

brain structure processing information about smells, demonstrated drift of odorant representations despite stable odor identification. Such drift was characterized by a gradual and across-days decay in the self-similarity of the stimulus representation. Additionally, in the same study, the drift was shown to be stimulus dependent. It was shown that a more familiar stimulus drifts at a smaller rate (Schoonover et al., 2021).

The mechanisms and implications of the drift in the brain and artificial neural networks are still under investigation (see recent reviews by Masset et al. (2022), Rule et al. (2019), and Driscoll et al. (2022)). Recent modeling studies have shown that drift could happen in the presence of synaptic or other types of noise (Qin et al., 2023; Aitken et al., 2022). Using simulations of neural networks, Aitken et al. (2022) showed different noise types injected during training could lead to drift with qualitatively different patterns and geometries. However, no theoretical consideration of the drift was provided in that study. Qin et al. (2023) studied the drift in a network with a similarity-matching objective and a Hebbian/anti-Hebbian learning rule. They showed that noisy synaptic updates could lead to a random-walk exploration of the solution space. Other mechanisms have been suggested on how a stable readout could be performed despite an evolving population code (Rule & O'Leary, 2022; Rule et al., 2020; Kalle Kossio et al., 2021).

One important source of noise during learning for both natural and artificial neural networks is the sampling noise that arises from the stochasticity in observing the data. It is therefore natural to ask if and how this type of noise can lead to drift, and whether it can explain the stimulus-dependency of the rate of the drift observed in experiments. Here, we aim to answer these questions in a two-layer linear neural network model that undergoes online continual learning via Stochastic Gradient Descent (SGD). In addition to being common in machine learning, feed-forward networks are a reasonable first approximation to sensory (olfactory) processing in the brain. Specifically, we found the two-layer linear network to be one of the simplest models that enables studying the representational drift in the hidden layer, and yet allows for analytical tractability.
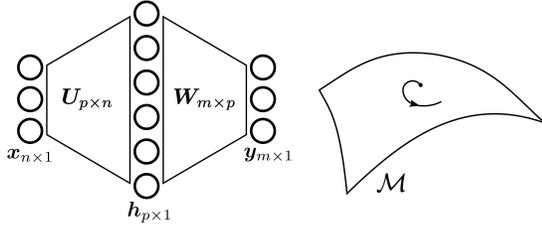
Figure 1. (left) Neural network model. (right) Manifold of minimum-loss ($\mathcal{M}$) in the parameter space.

2. Model and Theory

A multilayer network including L layers can be described by a set of weight matrices $\boldsymbol{W}^{(l)}$, where index l enumerates the individual layers. The network can be represented by a vector in a vector space constructed from non-commuting weight matrices:

$$\boldsymbol{\theta} = \left(\boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, ..., \boldsymbol{W}^{(L)} \right). \quad (1)$$

Our goal is to study the dynamics of learning and the drift in this space for a two-layer neural network described below (Section 2.1). We derive the manifold of stable minimum-loss (Section 2.2), and study its first order differential geometry (Section 2.3). We use the Euclidean inner product in this space, which means for two vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, we have:

$$\boldsymbol{\theta}_1^T \boldsymbol{\theta}_2 = \sum_l \mathrm{tr}\left(\boldsymbol{W}_1^{(l)T} \bar{\boldsymbol{W}}_2^{(l)} \right). \quad (2)$$

By characterizing the movements normal and tangential to the manifold (Sections 2.4.1 and 2.4.2), we derive an effective diffusion process on the manifold and calculate the corresponding diffusion ("drift") rates for the representations (Section 2.4.3). Finally, we apply this framework to study the dependency of the drift on the input statistics in two cases of isotropic Gaussian stimuli (Section 3), and a case with a frequently presented stimulus (Section 4). Our results are further validated by numerical simulations.

2.1. Neural network model

Our model consists of a linear feedforward neural network with an expansive hidden layer, as shown in Figure 1. The input to the network is the external stimulus ($\boldsymbol{x} \in \mathbb{R}^n$), the hidden layer activation is the representation of the stimulus ($\boldsymbol{h} \in \mathbb{R}^p$, e.g. neural activities in the piriform cortex), and the output is the task outcome ($\boldsymbol{y} \in \mathbb{R}^m$, e.g. the percept or an identity of an stimulus). The two weight matrices are $\boldsymbol{U} = \boldsymbol{W}^{(1)}$ and $\boldsymbol{W} = \boldsymbol{W}^{(2)}$ respectively, and the predicted output is $\hat{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{U}\boldsymbol{x}$. In the parameter space the network is denoted by $\boldsymbol{\theta} = (\boldsymbol{U}, \boldsymbol{W})$.

We consider a continual online learning scenario in which the network sees one sample at a time taken independently from a fixed data distribution. The objective function consist of a Mean Squared Error (MSE) loss and L2-regularization. Hence, the sample loss becomes:

$$l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{W}\boldsymbol{U}\boldsymbol{x}\|^2 + \frac{\gamma}{2}(\|\boldsymbol{W}\|_F^2 + \|\boldsymbol{U}\|_F^2), \quad (3)$$

where γ is the regularization coefficient. Further, we assume $\boldsymbol{y} = \boldsymbol{x}$, which means the goal of the network is to learn the identity mapping from the input to the output. Note this essentially becomes an autoencoder with an expansive hidden layer ($p \geq n = m$). Finally, the learning occurs via the SGD with a minibatch size of one. The update in the parameter vector $\boldsymbol{\theta}$ upon observing sample $\boldsymbol{x}$ is:

$$\Delta \boldsymbol{\theta} = -\eta \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}), \quad (4)$$

where $\boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} l(\boldsymbol{x}; \boldsymbol{\theta})$ is the sample gradient vector, and η is the learning rate. In terms of the weight matrices, this is equivalent to $\Delta \boldsymbol{W} = -\eta \nabla_{\boldsymbol{W}} l$ and $\Delta \boldsymbol{U} = -\eta \nabla_{\boldsymbol{U}} l$.

2.2. Degeneracy and the manifold of solutions

A condition that makes representational drift possible is the redundancy of the network parameters associated with achieving a given task performance. In a two-layer network with $\hat{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{U}\boldsymbol{x}$, the transformations $\tilde{\boldsymbol{W}} \to \tilde{\boldsymbol{W}}\boldsymbol{Q}$ and $\tilde{\boldsymbol{U}} \to \boldsymbol{Q}^{-1}\tilde{\boldsymbol{U}}$ leave the output unchanged for any invertible matrix $\boldsymbol{Q}$. In our model, the presence of L2 regularization confines $\boldsymbol{Q}$ to be orthogonal for the loss to remain unchanged. The effect of $\boldsymbol{Q}$ is rotation of the representations, and hence it links the degeneracy to a rotational symmetry in the network. We aim to study whether and how the space of redundant parameters are explored due to the online learning stochasticity and, from that, characterize the rate and the geometry of the drift in the representations.

We define the *manifold of solutions*, $\mathcal{M}$, to be the set of stable critical points in the parameter space. This manifold represents the redundancy in the model, as all the points on it have the same expected loss value $L(\boldsymbol{\theta}) = \langle l(\boldsymbol{\theta}) \rangle_x$, and hence are equally preferable (note that $\langle . \rangle_x$ is the expectation over the input distribution). We analytically derive $\mathcal{M}$ in the following theorem as a function of the input distribution, and for the rest of the paper refer to it as the *manifold*.

Theorem 2.1. (Manifold) *The manifold of solutions for learning the identity map ($\boldsymbol{y} = \boldsymbol{x}$) is:*

$$\mathcal{M} : \{\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{W}}) \,|\, \tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^T = \boldsymbol{I}_n - \gamma \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}, \tilde{\boldsymbol{U}} = \tilde{\boldsymbol{W}}^T\}, \quad (5)$$

where $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \langle \boldsymbol{x}\boldsymbol{x}^T \rangle_x$ is the second order moment of the input distribution with eigenvalues that are assumed to be greater than γ.

The above can be shown by first finding the critical points of the loss satisfying $\nabla_{\boldsymbol{U}} L = \nabla_{\boldsymbol{W}} L = 0$. The solutions consist of saddle-points, as well as the above manifold which

has a non-negative Hessian eigenspectrum (see the proof in Appendix A).

2.3. Differential geometry of the manifold

In order to study the learning dynamics near the manifold, we characterize its first order differential geometry. This is done by finding the local tangent and normal spaces to the manifold ($T_{\tilde{\theta}}\mathcal{M}$ and $N_{\tilde{\theta}}\mathcal{M}$, respectively).

Lemma 2.2. (Tangent space) *The local tangent space to the manifold at point $(\tilde{W}^T, \tilde{W})$ is spanned by vectors $t = (T_W^T, T_W)$ with $T_W = \tilde{W}\tilde{W}^T\Omega\tilde{W} + K\tilde{W}_\perp$, where Ω is an arbitrary $n \times n$ skew-symmetric matrix, K is an arbitrary $n \times (p-n)$ matrix, and $\tilde{W}_\perp$ is a full-rank $(p-n) \times p$ matrix whose rows are orthogonal to rows of $\tilde{W}$.*

The proof is presented in Appendix B. To gain some intuitions about the above results, note that movements in the tangent space create coordinated changes $\Delta W = T_W$ and $\Delta U = T_W^T$ in the weight matrices such that their effects cancel out at the output (this is because $\Delta(\tilde{W}\tilde{U}) \approx \tilde{W}T_W^T + T_W\tilde{U} = 0$). Additionally, these changes generate rotations of representation vectors, which is expected due to the rotational symmetry of the problem discussed in the previous section. In this context, the term in T_W containing Ω creates rotations within the n-dimensional subspace of the existing representations (i.e. the column-space of $\tilde{U}$), while the term containing K is responsible for rotations that move the representations toward the $(p-n)$-dimensional subspace orthogonal to the current representations[1].

Using the inner product in Eq.2, the normal space can be defined as the subspace orthogonal to the tangent space. We will use the eigenspace of the Hessian corresponding to positive eigenvalues to describe the normal space (see Section 2.4 and Appendix A.2.1 – note similarly, the tangent space is equivalent to the eigenspace of the Hessian with zero eigenvalue). In the subsequent sections, we use the results in this section to project the gradient vector into these spaces.

2.4. Learning dynamics near the manifold and emergence of drift

In this section, we will study the stochastic dynamics of learning in an online learning scenario in which the data are sampled from a fixed distribution. We assume sufficient training has been performed, and hence we are on or close to the manifold while being continuously nudged around due to the SGD noise. An arbitrary point θ near the manifold can be represented as:

$$\theta = \tilde{\theta} + \theta_N, \tag{6}$$

where $\tilde{\theta}$ is the closest point on the manifold to θ, and $\theta_N \in N_{\tilde{\theta}}\mathcal{M}$ is the deviation from the manifold in the normal space. In the rest of this section, we will find update equations for θ_N and $\tilde{\theta}$, and describe their dynamics.

2.4.1. FLUCTUATION IN THE NORMAL SPACE

We can find an update equation for θ_N by projecting the two sides of the SGD update (Eq.4) into the normal space. For small learning rates and over long times, this equation can be approximated with a continuous-time stochastic differential equation (SDE):

$$d\theta_N = -H\theta_N dt + \sqrt{\eta}\, C d B_t. \tag{7}$$

In the above, $H_{k,l\,(\in[1,2np])} = \frac{\partial L}{\partial \theta_k \partial \theta_l}\big|_{\tilde{\theta}}$ is the Hessian, B_t is the standard multi-dimensional Brownian motion (Wiener process), and $C \approx \langle \tilde{g}\tilde{g}^T\rangle_x^{1/2}$, where $\tilde{g}$ is the gradient at point $\tilde{\theta}$ on the manifold (Eq.8). In general, the gradient (g) can be analytically calculated at any point by differentiating the loss function with respect to the weight matrices (see Eq.31 in the Appendix). On the manifold this becomes:

$$\tilde{g}(x) := g(x; \tilde{\theta}) = (\tilde{W}^T Z_x, Z_x \tilde{W}), \tag{8}$$

where $Z_x = \gamma(I_n - \Sigma_x^{-1}xx^T)$. Note that C in Eq.7 is in general a function of θ_N. But in deriving the SDE, we approximated it with its value on the manifold ($\theta_N = 0$), which is justified for small deviations (see Section C.1 for the derivation of the SDE).

Since the Hessian is positive semidefinite on the manifold, the process defined by the SDE in Eq.7 is a mean-reverting process known as the multi-dimensional Ornstein-Uhlenbeck (OU) stochastic process (Gardiner et al. (1985)). OU has a stationary solution with zero mean and a finite variance (Appendix C.2). Hence, we refer to the deviations in the normal space as *fluctuation*. We represent the fluctuation in a basis constructed from the eigenvectors of the Hessian with positive eigenvalues:

$$\rho = N^T\theta_N, \quad \text{where } N = [n_1|n_2|...|n_K], \tag{9}$$
$$\text{and } Hn_k = \lambda_k n_k, \ \lambda_k > 0.$$

Here ρ_k represents the deviation from the manifold along the Hessian eigenvector n_k. As shown in Appendix C.2, the stationary solution of ρ satisfies $\langle \rho_k \rangle = 0$, and has the covariance:

$$\langle \rho_k \rho_l \rangle = \frac{\eta}{\lambda_k + \lambda_l}\langle n_k^T\tilde{g}(x)\, n_l^T\tilde{g}(x)\rangle_x. \tag{10}$$

As expected, since $\tilde{g}(x)$ is the driver of the fluctuation in the SDE (Eq.7), the fluctuation covariance depends on the projection of $\tilde{g}$ on $\{n_k\}$. In Appendix A.2.1 we calculate these eigenvectors, which consists of two sets. We show that $\tilde{g}$ has no projection on one of these sets and hence, for

---

[1]Note the dimension of the tangent space is $np - \frac{n(n+1)}{2}$, which is equal to the degrees of freedom in Ω and K.

the purpose of fluctuation, only a subspace of the normal space is relevant. This subspace is described in the next proposition.

Proposition 2.3. (Hessian eigenspace) *If (v_i, s_i) are the eigenvector/eigenvalue pairs of $\Sigma_x = \langle xx^T \rangle$ with $s_1 \geqslant .. \geqslant s_n > \gamma$, the Hessian eigenvectors n_k along which there is non-zero fluctuation (i.e. $n_k^T \tilde{g} \neq 0$) correspond to (v_i, v_j) pairs for $i, j \in [1, n]$ via:*

$$n_k \equiv n_{ij} = (\tilde{W}^T Z_{ij}, Z_{ij} \tilde{W}), \qquad (11)$$

where $Z_{ij} = C_{ij}(\kappa_{ij} v_i v_j^T + v_j v_i^T)$. Here, $\kappa_{ij} = sgn(i - j)(\sqrt{1 + b^2} - b)$ for $b = [1/\gamma - (s_i + s_j)/(2 s_i s_j)]|s_i - s_j|$, and $C_{ij} = [(2 - \gamma/s_i - \gamma/s_j)(1 + \kappa_{ij}^2)]^{-\frac{1}{2}}$ is a normalization constant ensuring $n_k^T n_k = 1$. The corresponding Hessian eigenvalues are $\lambda_{ii} = 2(s_i - \gamma)$ and $\lambda_{ij(i \neq j)} = 2 s_i - \gamma(s_i/s_j + \kappa_{ij})$. (See proof in Appendix C.3.)

The above proposition relates the relevant directions in the normal space to the corresponding (ij) indices of the input principal directions (this makes k a composite index). Subsequently, the components of the fluctuation covariance can be written as $\langle \rho_k \rho_l \rangle \equiv \langle \rho_{ij} \rho_{pq} \rangle$ (see Appendix C.4 for the components as functions of the input distribution). The remark below provides an interpretation of movements along different n_k in the representation space.

Remark 2.4. (Fluctuation of the representations): When on the manifold, the representations of all unit-length stimuli form an n-dimensional ellipsoid embedded in $\mathbb{R}^p$. The main axes of this ellipsoid are $\tilde{h}_i = \tilde{W}^T v_i$, with norms $|\tilde{h}_i| = (1 - \gamma/s_i)^{1/2}$ ($i \in [1, n]$). Using the above proposition, we can determine how deviations along different n_k lead to different deformation modes for the ellipsoid. Specifically, the deviation along n_{ii} changes $\tilde{h}_i$ radially (changing its norm), while leaving other (orthogonal) axes unchanged. Similarly, deviation along n_{ij} for $i \neq j$ moves $\tilde{h}_i$ and $\tilde{h}_j$ toward or away from each other, and leaves other axes intact. The latter shows that the fluctuations can change not only the lengths but also the pairwise angles of the representation vectors. Note however that the average of these changes is zero over time.

We take the variance of the norm as a measure of the fluctuation for the representations, i.e.

$$\sigma_i^2 := \text{var}(|h_i(t)|) = \frac{1}{2} \langle \rho_{ii}^2 \rangle = \frac{\eta \gamma^2}{4 s_i} \left(\frac{\langle x_i^4 \rangle_x}{s_i^2} - 1 \right), \quad (12)$$

where $x_i := v_i^T x$ (see Section C.5 for derivation). As we show later, this has an excellent match to numerically measured fluctuations.

2.4.2. TANGENTIAL UPDATES

Over a learning update, the displacement in the parameter space ($\Delta \theta$) leads to a corresponding projected movement on

the manifold ($\Delta \tilde{\theta}$). For small learning rates and by avoiding the curvature effect, we have:

$$\Delta \tilde{\theta} = -\eta g_T(x; \tilde{\theta} + N\rho), \qquad (13)$$

where $g_T = \Pi_T(g)$ is the projection of the gradient vector onto the tangent space. In the next theorem, we calculate this projection and find its action on the representations.

Theorem 2.5. (Tangential update) *For a normal deviation $N\rho = \sum_k \rho_k n_k$ from point $\tilde{\theta} = (\tilde{W}^T, \tilde{W})$ on the manifold, the tangential projection of the gradient is:*

$$g_T(x; \tilde{\theta} + N\rho) = (G_{U_T}, G_{U_T}^T), \quad for \qquad (14)$$
$$G_{U_T} = \tilde{W}^T (\tilde{W}\tilde{W}^T)^{-\frac{1}{2}} \left(\sum_k \rho_k \mathcal{G}_k^{:,:} \right) (\tilde{W}\tilde{W}^T)^{\frac{1}{2}} + \mathcal{O}(|\rho|^2),$$

where $\mathcal{G}$ is a rank-3 tensor with components $\mathcal{G}_k^{s,r} \equiv \mathcal{G}_{ij}^{s,r}$ that are defined as:

$$\mathcal{G}_{ij}^{s,r}(x) = \frac{\gamma C_{ij} \sqrt{\omega_r \omega_s}}{2(\omega_r + \omega_s)} [x_j (S_{sj}^{ij} x_s \delta_i^r - S_{rj}^{ij} x_r \delta_i^s)$$
$$+ x_i (S_{is}^{ij} x_s \delta_j^r - S_{ir}^{ij} x_r \delta_j^s)], \quad (15)$$

where δ_s^r is the Kronecker delta function, $\omega_i := 1 - \gamma/s_i$, and $S_{rs}^{ij} := \kappa_{ij}/s_r + 1/s_s$ ($r, s, i, j \in [1, n]$). The action of the tangential gradient in the representation space is a small rigid-body rotation around the origin. In this rotation, the representation vectors stay within the column-space of $\tilde{W}^T$, which itself stays fixed during the tangential updates. Additionally, the angular displacement of the representation $h_s(= Uv_s)$ toward $h_r(= Uv_r)$ is:

$$\Delta \varphi_{sr}(x; \tilde{\theta} + N\rho) = \eta \sum_{k=1}^K \rho_k \mathcal{G}_k^{s,r}(x) + \mathcal{O}(|\rho|^2). \quad (16)$$

In deriving the above, the gradient was approximated to the first order of the deviation from the manifold, and subsequently the tangential projector operator was applied. Here $\mathcal{G}$ is a tensor that, for a given direction in the normal space (corresponding to $k \equiv ij$), returns a rotation generator via the skew-symmetric matrix $\mathcal{G}_k^{:,:}$ ($G_k^{s,r} = -G_k^{r,s}$). Note also that during the gradient updates only a subspace of the tangent space is explored corresponding to the term with Ω in Lemma 2.2 (see Appendix D).

2.4.3. DRIFT AS A DIFFUSION PROCESS

As we saw in Section 2.4.1, the mean-reverting property of the gradient confines the normal deviations near the manifold (Fig.2 - Left). The movements in the tangent space, however, face no such mean-reverting property and hence can diffuse freely on the manifold in a random-walk fashion. As per Theorem 2.5, the tangential component of the gradient is proportional to the deviation from the manifold (ρ),
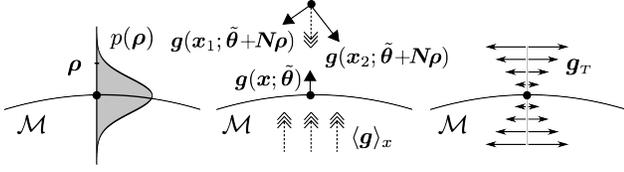
Figure 2. Schematics showing (left) the probability distribution of the fluctuations outside the manifold, (middle) the gradient vectors on and near the manifold, and (right) the tangential component of the gradient.

and it vanishes on the manifold (see the schematics in the middle and right panels of Fig.2). We can find an effective diffusion process for the movements on the manifold, which because of the above correspondence, is expected to depend on the statistics of the normal deviations. Specifically, over large timescales, Eq.13 can be approximated as the following continuous-time SDE, describing the evolution of the the parameter vector on the manifold:

$$d\tilde{\boldsymbol{\theta}} = \sqrt{2\eta^{-1}\boldsymbol{D}_\theta}\,d\boldsymbol{B}_t. \tag{17}$$

In the above, $\boldsymbol{D}_\theta = (1/2)\langle\Delta\tilde{\boldsymbol{\theta}}\Delta\tilde{\boldsymbol{\theta}}^T\rangle_{x,\rho} = (\eta^2/2)\langle\boldsymbol{g}_T\boldsymbol{g}_T^T\rangle_{x,\rho}$ is the diffusion tensor for the parameters (measured over a training step), and $\boldsymbol{B}_t$ is multi-dimensional Brownian motion. The above diffusion process is manifested as a rotational diffusion in the representation space. According to Theorem 2.5, the rotations happen within the n-dimensional space of representations and hence to fully characterize the diffusion it's sufficient to measure the pairwise diffusion rates between the axes of the ellipsoid $\{\boldsymbol{h}_s\}$ ($s \in [1, n]$). Similar to the parameter diffusion above, we define D_{sr} to be the diffusion rate between the two representations $\boldsymbol{h}_s$ and $\boldsymbol{h}_r$ based on the mean squared of the pairwise angular displacement, i.e.:

$$D_{sr} := \frac{1}{2}\langle\Delta\varphi_{sr}^2\rangle_{x,\rho} = \frac{\eta^2}{2}\sum_{k,l=1}^{K}\langle\rho_k\rho_l\rangle\bar{\bar{\boldsymbol{\mathcal{G}}}}_{k,l}^{s,r}, \tag{18}$$

where we defined $\bar{\bar{\boldsymbol{\mathcal{G}}}}_{k,l}^{s,r} := \langle\boldsymbol{\mathcal{G}}_k^{s,r}\boldsymbol{\mathcal{G}}_l^{s,r}\rangle_x$. The right-hand side results from replacing $\Delta\varphi_{sr}$ from Eq.16 and taking the average. The total diffusion for the representation $\boldsymbol{h}_s$ can be derived by summing up the diffusion rates along different directions: $D_s = \sum_{r=1,r\neq s}^{n}D_{sr}$.

Eq.18 indicates that the diffusion rate between two representation vectors is an aggregation of terms resulting from deviations in different directions in the normal space. In this equation, $\eta^2\bar{\bar{\boldsymbol{\mathcal{G}}}}_{k,l}$ can be thought of as the diffusion per direction (or more accurately direction pair k, l) in the normal space, and $\langle\rho_k\rho_l\rangle$ is the covariance of the corresponding fluctuations. As both $\bar{\bar{\boldsymbol{\mathcal{G}}}}$ and $\langle\boldsymbol{\rho}\boldsymbol{\rho}^T\rangle$ depend on the second and the fourth moments of the stimuli distribution, the diffusion is also a function of the those moments of the input distribution.

2.5. Numerical simulations

Alongside the analytical derivations, we also performed numerical simulations in which we measured the drift in a neural network undergoing continual SGD training. We first numerically validated the equations of the manifold by verifying that after enough training steps the network approaches the theoretical manifold. Next, to measure the drift, we initialized multiple ($> 10^4$) realizations of the network, all starting from a fixed point on the manifold but undergoing different random SGD samplings during the training. Following a transitory phase, we studied the over-time trajectories of the hidden layer activation for different trial stimuli. This was done by measuring the fluctuation (variance) of the representation norm, and the angular displacements of the representations. The diffusion coefficient was estimated as half of a linear fit slope to the mean squared angular displacements aggregated from all the realizations.

3. Drift Under Gaussian Stimuli

In this section, we present complete analytical results for a case where the stimuli are drawn randomly and independently from a standard n-dimensional Gaussian distribution, i.e. $x_i \sim \mathcal{N}(0, 1)$. Since the input covariance is the identity ($\langle x_i x_j\rangle = \delta_i^j$), its eigenvectors $\{\boldsymbol{v}_i\}$ form an orthonormal basis for $\mathbb{R}^n$, all corresponding to eigenvalue $s_i = 1$. Using Proposition 2.3, the relevant directions in the normal space consist of three sets corresponding to $\boldsymbol{Z}_{ii} = \boldsymbol{v}_i\boldsymbol{v}_i^T/(\sqrt{2\omega})$, $\boldsymbol{Z}_{ij(i>j)} = (\boldsymbol{v}_i\boldsymbol{v}_j^T + \boldsymbol{v}_j\boldsymbol{v}_i^T)/(2\sqrt{\omega})$, and $\boldsymbol{Z}_{ij(i<j)} = (-\boldsymbol{v}_i\boldsymbol{v}_j^T + \boldsymbol{v}_j\boldsymbol{v}_i^T)/(2\sqrt{\omega})$, where $\omega = 1 - \gamma$, and $i, j \in [1, n]$. For calculating the above, we derived the coefficients $\kappa_{ij(i>j)} = 1$, $\kappa_{ii} = 0$, and $\kappa_{ij(i<j)} = -1$ in the proposition. The corresponding Hessian eigenvalues are $\lambda_{ij(i\geqslant j)} = 2(1 - \gamma)$ and $\lambda_{ij(i<j)} = 2 - \gamma$.

The components of the fluctuation matrix result from projecting the gradient onto the above eigenvectors (Eq.10) and using the identity $\langle x_i x_j x_p x_q\rangle = \delta_{ip}^{jq} + \delta_{ij}^{pq} + \delta_{ij}^{qp}$ for the Gaussian input (see Appendix C.4). This leads to:

$$\langle\rho_{ij}^2\rangle = \eta\gamma^2, \text{ for } i \geqslant j, \tag{19}$$

with the rest of the components being zero. Following Eq.12, the fluctuation in the representation norm of an arbitrary unit-length stimulus becomes:

$$\sigma_s^2 = \frac{\eta\gamma^2}{2}. \tag{20}$$

To find the diffusion coefficients, we first calculate the components of the $\boldsymbol{\mathcal{G}}$ tensor from Eq.15:

$$\boldsymbol{\mathcal{G}}_{ii}^{s,r} = \frac{\gamma}{2\sqrt{2\omega}}x_i(x_s\delta_i^r - x_r\delta_i^s), \quad \boldsymbol{\mathcal{G}}_{ij(i<j)}^{s,r} = 0, \tag{21}$$

$$\boldsymbol{\mathcal{G}}_{ij(i>j)}^{s,r} = \frac{\gamma}{4\sqrt{\omega}}[x_j(x_s\delta_i^r - x_r\delta_i^s) + x_i(x_s\delta_j^r - x_r\delta_j^s)].$$

5

Subsequently, the coefficients $\bar{\bar{\mathcal{G}}}^{s,r}_{ij,pq}$ can be found by taking the averages of the products of the $\mathcal{G}$ terms from Eq.21, leading to:

$$\bar{\bar{\mathcal{G}}}^{s,r}_{ii,ii} = \frac{\gamma^2}{8\omega}(\delta^r_i + \delta^s_i), \quad \bar{\bar{\mathcal{G}}}^{s,r}_{ii,jj(i\neq j)} = \frac{-\gamma^2}{8\omega}(\delta^{rs}_{ij} + \delta^{sr}_{ij}),$$

$$\bar{\bar{\mathcal{G}}}^{s,r}_{ij,ij(i>j)} = \frac{\gamma^2}{16\omega}(\delta^r_i + \delta^s_i + \delta^r_j + \delta^s_j + 2\delta^{rs}_{ij} + 2\delta^{sr}_{ij}), \quad (22)$$

with the rest of the components being zero. Finally, replacing the above in the summation of Eq.18 results in the diffusion rate between the representations of two arbitrary stimuli denoted by s and $r(\neq s)$:

$$D_{sr} = \frac{\eta^3\gamma^2}{2}\left(\sum_{i=1}^{n}\bar{\bar{\mathcal{G}}}^{s,r}_{ii,ii} + \sum_{i,j=1,i>j}^{n}\bar{\bar{\mathcal{G}}}^{s,r}_{ij,ij}\right)$$

$$= \frac{\eta^3\gamma^4}{2(1-\gamma)}\left(\frac{1}{4} + \frac{n}{8}\right) = \frac{1}{16}\frac{\eta^3\gamma^4}{1-\gamma}(n+2). \quad (23)$$

Note the summation was performed only over indices where the fluctuation covariance was non-zero. Since the diffusion is isotropic, the total diffusion for representation $\tilde{h}_s$ becomes:

$$D_s = (n-1)D_{sr} = \frac{1}{16}\frac{\eta^3\gamma^4}{1-\gamma}(n-1)(n+2). \quad (24)$$

4. Drift Under a Frequent Stimulus

As demonstrated in the previous section, SGD-induced drift could occur even with isotropic background stimuli. In this section, we will study how the presence of a frequent stimulus in the environment influences the drift. To do this, we consider a case where in addition to the background Gaussian stimuli, there is a relatively more frequent stimulus, $\boldsymbol{a}$, that is presented with probability α during training, i.e.:

$$\boldsymbol{x} = \begin{cases} \boldsymbol{a} & \Pr = \alpha \\ \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n) & \Pr = 1-\alpha \end{cases} \quad (25)$$

Note that the previous case in Section 3 is equivalent to $\alpha = 0$, and here we study the drift as a function of α. Without loss of generality, we take $\boldsymbol{a}$ to be along the first axis, and for simplicity we assume it has unit length. The second and fourth moments of the input distribution become $\langle x_i x_j \rangle = \alpha\delta^{11}_{ij} + (1-\alpha)\delta^j_i$ and $\langle x_i x_j x_p x_q \rangle = \alpha\delta^{1111}_{ijpq} + (1-\alpha)(\delta^{jq}_{ip} + \delta^{pq}_{ij} + \delta^{qp}_{ij})$, respectively. The eigenvectors of $\boldsymbol{\Sigma_x}$ still form an orthonormal basis for the input, but with eigenvalues $s_a = s_1 = 1$, and $s_b = 1-\alpha$ for $b \in [2, n]$ (note we use indices a and b to refer to the frequent and orthogonal background stimuli, respectively). The fluctuations of the representation norm for the frequent stimulus $\boldsymbol{a}$ and a unit length background stimulus $\boldsymbol{b}\,(\perp \boldsymbol{a})$ can be found from Eq.12:

$$\sigma^2_a = \frac{\eta\gamma^2}{2}(1-\alpha), \quad \sigma^2_b = \frac{\eta\gamma^2}{4}\frac{2+\alpha}{(1-\alpha)^2}. \quad (26)$$

It is easy to check that $\sigma^2_a \leqslant \sigma^2_b$ irrespective of α, which suggests that the fluctuation of the frequent stimulus representation is smaller than that of a background stimulus.

To fully characterize the diffusion in the representation space, we take advantage of the symmetry within the space of background stimuli. This means we only need to find two pairwise diffusion coefficients: D_{ab} $(= D_{ba})$, which measures the diffusion between the frequent and a background stimulus, and D_{bc} $(= D_{cb})$, which measures the pairwise diffusion between two orthogonal background stimuli (without loss of generality, we take $b = 2$ and $c = 3$). To calculate these coefficients, we need to perform the summation in Eq.18 which is taken over indices (k, l) or equivalently (ij, pq) for $i, j, p, q \in [1, n]$. For large n, many of the terms in the summation can be ignored as shown in Appendix E.1. Here we present results for two limiting cases.

We first study the regime of small α, which can be considered as a perturbation to the previous Gaussian stimuli case. As a first order correction to the summation, it is sufficient to sum over indices for which the unperturbed case had non-zero $\langle \rho_{ij}\rho_{pq} \rangle$ or $\bar{\bar{\mathcal{G}}}^{s,r}$ terms. This limits the summation to indices $(ij, ij)_{i>j}$ for which either i or j are equal to r or s (see details in Appendix E.1). If we take d $(> c)$ to be an additional index within the background subspace (e.g. $d = 4$), the diffusion summations for the two coefficients simplify to:

$$D_{ba} \approx \frac{n\eta^2}{2}(\langle\rho^2_{db}\rangle\bar{\bar{\mathcal{G}}}^{a,b}_{db,db} + \langle\rho^2_{da}\rangle\bar{\bar{\mathcal{G}}}^{a,b}_{da,da}) \quad (\alpha \ll 1, n \gg 1)$$

$$\approx \frac{n\eta^3\gamma^4}{32}[(1+2\alpha)(1) + (1+\frac{\alpha}{2})(1+\alpha)]$$

$$\approx \frac{n\eta^3\gamma^4}{16}(1 + \frac{7\alpha}{4})$$

$$D_{bc} \approx \frac{n\eta^2}{2}(\langle\rho^2_{db}\rangle\bar{\bar{\mathcal{G}}}^{c,b}_{db,db} + \langle\rho^2_{dc}\rangle\bar{\bar{\mathcal{G}}}^{c,b}_{dc,dc})$$

$$\approx \frac{n\eta^3\gamma^4}{32}[(1+2\alpha)(1+\alpha) + (1+2\alpha)(1+\alpha)]$$

$$\approx \frac{n\eta^3\gamma^4}{16}(1 + 3\alpha), \quad (27)$$

where we replaced the following quantities calculated in Appendix E.2 under small α and γ:

$$\langle\rho^2_{db}\rangle = \langle\rho^2_{dc}\rangle = \eta\gamma^2(1+2\alpha), \quad \langle\rho^2_{da}\rangle = \eta\gamma^2(1+\frac{\alpha}{2}) \quad (28)$$

$$\bar{\bar{\mathcal{G}}}^{a,b}_{db,db} = \frac{\gamma^2}{16}, \quad \bar{\bar{\mathcal{G}}}^{a,b}_{da,da} = \bar{\bar{\mathcal{G}}}^{c,b}_{db,db} = \bar{\bar{\mathcal{G}}}^{c,b}_{dc,dc} = \frac{\gamma^2}{16}(1+\alpha).$$

The above shows $D_{bc} - D_{ba} \approx 5n\eta^3\gamma^4\alpha/64 \geq 0$. This difference can be attributed to both the fluctuation and the $\bar{\bar{\mathcal{G}}}$ terms (for example, one can check from Eq.27 that if the fluctuation terms are replaced with their values at $\alpha = 0$, the difference still remains but to a lesser degree).

Total diffusion for stimuli a and b becomes:

$$D_a = (n-1)D_{ab} \approx \frac{n^2\eta^3\gamma^4}{16}(1+\frac{7\alpha}{4}) \quad (\alpha \ll 1, n \gg 1)$$

$$D_b = D_{ba} + (n-2)D_{bc} \approx \frac{n^2\eta^3\gamma^4}{16}(1+3\alpha), \qquad (29)$$

which shows $D_b \geq D_a$.

In Appendix E.3, we also perform derivations for large n and $\alpha \gg \gamma$ (specifically, with respect to the input eigenvalues, we assume $s_a, s_b, s_a - s_b \gg \gamma$). The results of those calculations are:

$$D_a \approx \frac{n^2\eta^3\gamma^4\langle x_a^2 x_c^2\rangle\langle x_b^2 x_c^2\rangle}{128 s_b^5}[1 + 3\frac{s_b}{s_a} + 2(\frac{s_b}{s_a})^2 + \frac{4(\frac{s_b}{s_a})^2}{1+\frac{s_b}{s_a}}]$$

$$= \frac{n^2\eta^3\gamma^4}{16(1-\alpha)^3}[\frac{1-\frac{7}{4}\alpha+\frac{15}{16}\alpha^2-\frac{1}{8}\alpha^3}{1-\frac{\alpha}{2}}],$$

$$D_b \approx \frac{n^2\eta^3\gamma^4\langle x_b^2 x_c^2\rangle^2}{16 s_b^5} = \frac{n^2\eta^3\gamma^4}{16(1-\alpha)^3} \quad (\alpha \gg \gamma, n \gg 1)$$

$$(30)$$

where in the equalities we replaced $s_a = 1$, $s_b = 1 - \alpha$, and $\langle x_a^2 x_c^2\rangle = \langle x_b^2 x_c^2\rangle = 1 - \alpha$. The term inside the second line brackets in D_a is always smaller than one. Hence, we again have $D_b \geq D_a$.

In Figure 3, we plot the diffusion and fluctuations coefficients as a function of α for stimuli a and b. We see that, consistent with the above results, the frequent stimulus drifts at a relatively slower rate and has a smaller fluctuation, irrespective of α. Additionally, there is an excellent match between the theoretical and the simulations results for both the fluctuations and the diffusion coefficients. Finally, in the bottom panel of the same figure, we plotted the trajectories of the representations over time for $n = p = 3$. The lower fluctuation and diffusion rates for the more frequent stimulus can be visually observed from the smaller point cloud for this stimulus.

5. Discussion

In a two-layer neural network model of the olfactory system, we show, using theory and simulations, that the stochasticity in SGD online learning could result in a drift of stimuli representation over time, even after the training is complete and no measurable change in the performance is observed. We analytically demonstrate the dependency of the drift on the input distribution and, in particular, show that a frequently presented stimulus drifts at a relatively slower rate. This finding is consistent with experimental observations in the piriform cortex (Schoonover et al., 2021).

We studied the learning dynamics in the high-dimensional space of network parameters. In this space, drift can be
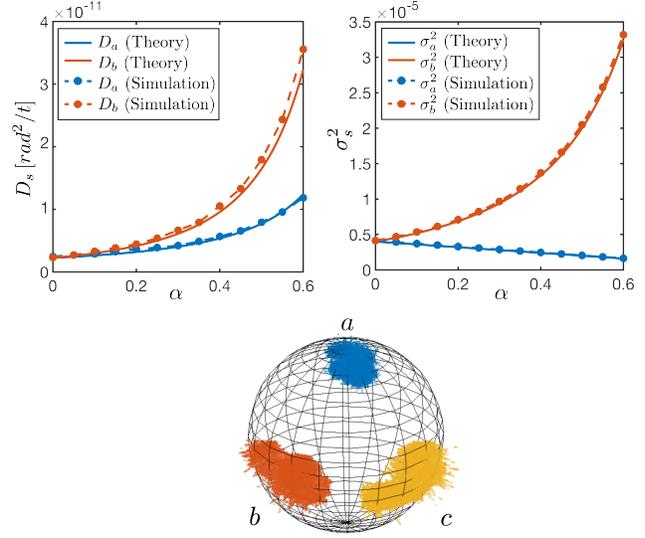
Figure 3. Plots of (left) diffusion, and (right) fluctuation for representations of a frequent (subscript a, blue) and a background stimulus (subscript b, red). Horizontal axes are the probability of the frequent stimulus, α. In both plots $m = n = 10, p = 20$, $\gamma = 0.04$, and $\eta = 0.005$. (bottom) History of representations for three trial stimuli after 2.2×10^5 training steps. $n = p = 3$, $\gamma = 0.1$, $\eta = 0.1$, and $\alpha = 0.5$. a is the frequent and b and c are two orthogonal background stimuli.

considered as any movement tangential to the manifold of minimum-loss, the effects of which aggregate over time to create an effective diffusion process. Orthogonal to this is the fluctuation outside the manifold that has a finite variance due to the mean-reverting property of the gradient. For the tangential gradient to have a non-zero value, an orthogonal deviation from the manifold was necessary (see Eq.14 and Figure 2). In a way, this makes the diffusion on the manifold a second order phenomenon, and that explains why the amount of diffusion depends on the fluctuation covariance (Eq.18). In our work, diffusion (random walk) emerged as a mechanism underlying the representational drift. Note that in this sense it is different from the term drift in Physics.

In the representation space, the effects of the fluctuation and tangential movements are deformations and rigid-body rotations of the space, respectively. If we consider the representation of a given stimulus over time, its trajectory consists of two parts: a random walk movement on a high-dimensional sphere, and simultaneously, a mean-reverting fluctuating process that causes deviations on and outside the sphere (i.e. changing norms and pair-wise angles of representations). This can be observed from the visualization in the bottom panel of Figure 3. Note that the random walk on a sphere dynamics is consistent with the over-time decay of population response self-similarity as observed in the experiments

of (Schoonover et al., 2021). Further, the lower diffusion rate observed for a more frequent stimulus suggests that the rigid-body rotation on average rotates the frequent stimulus to a lesser degree. This could be explained by, for example, the axis of rotation being on average closer to the representation of the frequent stimulus. The lower diffusion and fluctuation for the more frequent stimulus could be traced to the mechanism by which the gradient tends to preserve the representation and the output of more rehearsed stimuli..

Recently, Qin et al. (2023) studied drift in a one-layer neural network model of sensory systems with a similarity matching objective and a Hebbian/anti-Hebbian learning rule. Similarly to our results, they showed that the representations undergo a rotational diffusion process. This is not surprising as both models have objective functions with degenerate solutions and rotational symmetry. In our case, the L2 regularization creates that symmetry. Qin et al. (2023) also demonstrated a stimulus dependency of the diffusion coefficient, such that the drift for a given eigenvector direction was inversely related to its eigenvalue. However, in that study synaptic noise was injected to the network and the stimulus dependency was not studied when this noise was zero. Here, we used a two-layer neural network model with MSE loss to study drift under no external noise. We showed that the sampling noise due to the SGD stochasticity is enough to demonstrate the stimulus dependency of the drift rate. Along that line, we speculate that the anisotropic profile of the SGD noise enhances the stimulus dependency of the drift when compared to an isotropic (synaptic) noise. Our results further show that in the pure SGD case, the drift direction is limited to the subspace of existing stimuli representations. When an isotropic noise is injected to the weights, this will no longer be the case since the whole tangent space will be explored.

SGD dynamics have been studied extensively in the machine learning literature (see e.g. Mandt et al. (2017); Jastrzebski et al. (2017); Zhu et al. (2018); Chaudhari & Soatto (2018); Yaida (2018); Zhang et al. (2018)). In a recent study, Li et al. (2022) provided a mathematical framework for investigating the limiting dynamics of SGD around the local minimizer manifold in overparametrized networks. They used this framework to study the regularization effect of SGD (i.e. implicit bias), and showed that label noise could drive the network to areas with flatter landscape with potential benefits for generalization (Wei & Schwab, 2019; Blanc et al., 2020; Cowsik et al., 2022). Unlike that study, our analysis deals with a regime where the system has reached a final stage of learning with a steady local loss landscape. Hence in our case the effective dynamics on the manifold is a pure random-walk. Additionally, the analytical tractability of our model allowed us to calculate the exact gradient profile around the manifold. We found that to accurately calculate the drift, we had to integrate the effect of the tangential gradient not just on the manifold but away from it (hence the third power of the learning rate in our diffusion rate, one of which stems from the normal fluctuation covariance). Finally, our model contains weight decay, which influences both the learned manifold and the gradients.

Linear multilayer networks, despite their simplicity, demonstrate nonlinear and nontrivial learning behaviors (Saxe et al., 2013; Li & Sompolinsky, 2021). In our case, the two-layer feedforward network was the simplest architecture that provided necessary redundancy in the network parameters through the product of the weight matrices. We used the identity mapping (autoencoder) setup to simplify the derivations of main formalisms, allowing us to focus on the role of input distribution on the drift. The dependence of the drift on the data is inherent in the diffusion term in Eq.18, which ultimately depends on the $2^{nd}/4^{th}$ moments of the input distribution. Extensions of our work may include nonlinear and deep networks with recurrent connections in addition to a non-stationary data-distribution. The main concepts on which our framework builds are valid for more general linear or nonlinear networks minimizing an objective function with stationary data. In those cases, the loss could still be approximated as a quadratic function near the manifold of minimum loss, and hence the equations for fluctuations and tangent movements still hold. However, the specific form of the gradient as the driver of the drift and the geometry of the manifold will be problem-dependent. We expect our main findings (i.e. the presence of drift, and lower drift for a more frequent stimulus) to carry over to nonlinear cases, however, we leave that to future studies.

Acknowledgements

References

Absil, P.-A., Mahony, R., and Sepulchre, R. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Aitken, K., Garrett, M., Olsen, S., and Mihalas, S. The geometry of representational drift in natural and artificial neural networks. *PLOS Computational Biology*, 18(11): e1010716, 2022.

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pp. 483–513. PMLR, 2020.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles

for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018.

Cowsik, A., Can, T., and Glorioso, P. Flatter, faster: scaling momentum for optimal speedup of sgd. *arXiv preprint arXiv:2210.16400*, 2022.

Deitch, D., Rubin, A., and Ziv, Y. Representational drift in the mouse visual cortex. *Current Biology*, 31(19): 4327–4339, 2021.

Driscoll, L. N., Duncker, L., and Harvey, C. D. Representational drift: Emerging theories for continual learning and experimental future directions. *Current Opinion in Neurobiology*, 76:102609, 2022.

Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Gardiner, C. W. et al. *Handbook of stochastic methods*, volume 3. springer Berlin, 1985.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Kalle Kossio, Y. F., Goedeke, S., Klos, C., and Memmesheimer, R.-M. Drifting assemblies for persistent memory: Neuron transitions and unsupervised compensation. *Proceedings of the National Academy of Sciences*, 118(46):e2023832118, 2021.

Kunin, D., Bloom, J., Goeva, A., and Seed, C. Loss landscapes of regularized linear autoencoders. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3560–3569. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kunin19a.html.

Li, Q. and Sompolinsky, H. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Physical Review X*, 11(3):031059, 2021.

Li, Z., Wang, T., and Arora, S. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=siCt4xZn5Ve.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

Marks, T. D. and Goard, M. J. Stimulus-dependent representational drift in primary visual cortex. *Nature communications*, 12(1):1–16, 2021.

Masset, P., Qin, S., and Zavatone-Veth, J. A. Drifting neuronal representations: Bug or feature? *Biological cybernetics*, pp. 1–14, 2022.

Qin, S., Farashahi, S., Lipshutz, D., Sengupta, A. M., Chklovskii, D. B., and Pehlevan, C. Coordinated drift of receptive fields in hebbian/anti-hebbian network models during noisy representation learning. *Nature Neuroscience*, pp. 1–11, 2023.

Rule, M. E. and O'Leary, T. Self-healing codes: How stable neural populations can track continually reconfiguring neural representations. *Proceedings of the National Academy of Sciences*, 119(7):e2106692119, 2022.

Rule, M. E., O'Leary, T., and Harvey, C. D. Causes and consequences of representational drift. *Current opinion in neurobiology*, 58:141–147, 2019.

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

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Schoonover, C. E., Ohashi, S. N., Axel, R., and Fink, A. J. Representational drift in primary olfactory cortex. *Nature*, pp. 1–6, 2021.

Wei, M. and Schwab, D. J. How noise affects the hessian spectrum in overparameterized neural networks. *arXiv preprint arXiv:1910.00195*, 2019.

Yaida, S. Fluctuation-dissipation relations for stochastic gradient descent. *arXiv preprint arXiv:1810.00004*, 2018.

Zhang, Y., Saxe, A. M., Advani, M. S., and Lee, A. A. Energy–entropy competition and the effectiveness of stochastic gradient descent in machine learning. *Molecular Physics*, 116(21-22):3214–3223, 2018. doi: 10.1080/00268976.2018.1483535. URL https://doi.org/10.1080/00268976.2018.1483535.

Zhu, Z., Wu, J., Yu, B., Wu, L., and Ma, J. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

Ziv, Y., Burns, L. D., Cocker, E. D., Hamel, E. O., Ghosh, K. K., Kitch, L. J., El Gamal, A., and Schnitzer, M. J. Long-term dynamics of ca1 hippocampal place codes. *Nature neuroscience*, 16(3):264–266, 2013.

Appendix

A. Derivation of the Manifold and the Hessian (Proof of Theorem 2.1)

We lay out different parts of the proof over the next few sections. We first calculate all the critical points that satisfy the zero expected gradient condition (Section A.1), and from those, we calculate the Hessian (sections A.2 and A.2.1), and show that the manifold $\mathcal{M}$ is the only stable solution with non-negative Hessian spectrum (Section A.2.2). The proof is summarized in Section A.3.

A.1. Derivation of the critical points

The gradient can be found by taking the derivative of the loss function in Eq.3. By assuming $\boldsymbol{y} = \boldsymbol{x}$, we have:

$$\boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} l(\boldsymbol{x}; \boldsymbol{\theta}) = (\boldsymbol{G_U}, \boldsymbol{G_W}), \quad \begin{cases} \boldsymbol{G_W} = \nabla_W l = (\boldsymbol{WU} - \boldsymbol{I}_n)\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{U}^T + \gamma\boldsymbol{W} \\ \boldsymbol{G_U} = \nabla_U l = \boldsymbol{W}^T(\boldsymbol{WU} - \boldsymbol{I}_n)\boldsymbol{x}\boldsymbol{x}^T + \gamma\boldsymbol{U} \end{cases} \tag{31}$$

At the critical points the expected gradients are zero. This is equivalent to:

$$\langle \boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}) \rangle_x = \nabla_{\boldsymbol{\theta}} L = 0 \implies \begin{cases} (\boldsymbol{I}_n - \boldsymbol{WU})\boldsymbol{\Sigma_x}\boldsymbol{U}^T = \gamma\boldsymbol{W} \\ \boldsymbol{W}^T(\boldsymbol{I}_n - \boldsymbol{WU})\boldsymbol{\Sigma_x} = \gamma\boldsymbol{U} \end{cases} \tag{32}$$

which need to be solved for $\boldsymbol{W} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{U} \in \mathbb{R}^{p \times n}$ under $p \geq n$. A similar set of equations was studied in Kunin et al. (2019) but with $p \leq n$. We follow part of a proof from that study (Proof 2.1 in that paper) to show two facts about the solutions of the above equations (note that despite different assumptions for the dimension of the hidden layer, the following facts still hold).

Claim 1: $\boldsymbol{C} := (\boldsymbol{I}_n - \boldsymbol{WU})\boldsymbol{\Sigma_x} \succeq 0$. _Proof._ By multiplying the first line of Eq.32 by $\boldsymbol{W}^T$ from the right we get $(\boldsymbol{I}_n - \boldsymbol{WU})\boldsymbol{\Sigma_x}(\boldsymbol{WU})^T = \gamma\boldsymbol{WW}^T \succeq 0$, and therefore $\boldsymbol{\Sigma_x}(\boldsymbol{WU})^T \succeq (\boldsymbol{WU})\boldsymbol{\Sigma_x}(\boldsymbol{WU})^T$. Following a property of positive semi-definite matrices that states if $A \succeq 0$ and $AB^T \succeq BAB^T$, then $A \succeq BA$, we have $\boldsymbol{\Sigma_x} \succeq \boldsymbol{WU}\boldsymbol{\Sigma_x}$, which proves the claim. $\square$

Claim 2: $\boldsymbol{U} = \boldsymbol{W}^T$. _Proof._ By subtracting the transpose of the first line of Eq.32 from the second, and using the symmetric property of $\boldsymbol{C}$, we have: $(\boldsymbol{U} - \boldsymbol{W}^T)(\boldsymbol{C} + \gamma\boldsymbol{I}_n) = 0$. Now, since $\boldsymbol{C} + \gamma\boldsymbol{I}_n \succ 0$, we use a property of positive definite matrices that states if $B \succ 0$ and $A^TBA = 0$, then $A = 0$, to get $\boldsymbol{U} - \boldsymbol{W}^T = 0$. $\square$ For more information see Kunin et al. (2019) and its supplementary material.

Replacing $\boldsymbol{W}$ from the first line of Eq.32 into its second line, and using $\boldsymbol{U} = \boldsymbol{W}^T$ from Claim 2, we can form an equation for $\boldsymbol{WW}^T$:

$$(\boldsymbol{I}_n - \boldsymbol{WW}^T)\boldsymbol{\Sigma_x}(\boldsymbol{WW})^T(\boldsymbol{I}_n - \boldsymbol{WW}^T)\boldsymbol{\Sigma_x} = \gamma^2\boldsymbol{WW}^T \tag{33}$$

To solve the above we note that because we showed that $\boldsymbol{C} = (\boldsymbol{I}_n - \boldsymbol{WW}^T)\boldsymbol{\Sigma_x}$ is semi-positive definite and hence symmetric, two matrices $\boldsymbol{WW}^T$ and $\boldsymbol{\Sigma_x}$ commute and hence can be diagonalized simultaneously. Therefore, if $\boldsymbol{\Sigma_x} = \boldsymbol{VS_VV}^T$ is the singular value decomposition (SVD) with $\boldsymbol{S_V} = \text{diag}(\{s_i\})$, we can assume $\boldsymbol{WW}^T = \boldsymbol{V\Lambda V}^T$, where $\boldsymbol{\Lambda} = \text{diag}(\{\Lambda_i\})$ for $i \in [1, n]$. Replacing these in Eq.33, leads to n separate equations for the diagonal entries Λ_i:

$$\Lambda_i(\Lambda_i^2 s_i^2 - 2\Lambda_i s_i^2 + (s_i^2 - \gamma^2)) = 0, \quad i \in [1, n] \tag{34}$$

A valid solution for each entry is either $\Lambda_i = 0$ or $\Lambda_i = 1 - \frac{\gamma}{s_i}$ (note $\Lambda_i = 1 + \frac{\gamma}{s_i}$ also satisfies the above equation but it's not a valid solution overall since any $\boldsymbol{\Lambda}$ with such an entry violates the positive semi-definiteness of $\boldsymbol{C}$). For a given $\boldsymbol{\Lambda}$, we have $\boldsymbol{W} = \boldsymbol{V\Lambda}^{\frac{1}{2}}\boldsymbol{Q}^T$ and $\boldsymbol{U} = \boldsymbol{W}^T = \boldsymbol{Q\Lambda}^{\frac{1}{2}T}\boldsymbol{V}^T$ for any orthonormal matrix $\boldsymbol{Q}$. Based on the values of diagonal entries, we can classify the critical points into three sets:

- _The manifold ($\mathcal{M}$):_ $\Lambda_i = 1 - \frac{\gamma}{s_i}$ for all $i \in [1, n]$. In this case $\boldsymbol{\Lambda} = \boldsymbol{I}_n - \gamma\boldsymbol{S_V}^{-1}$ and hence $\boldsymbol{WW}^T = \boldsymbol{I}_n - \gamma\boldsymbol{\Sigma_x}^{-1}$.

- _Zero solution:_ $\Lambda_i = 0$ for all $i \in [1, n]$. In this case $\boldsymbol{W}$ and $\boldsymbol{U}$ are both zero.

- _Mixed solutions:_ $\Lambda_i = 0$ for some $i \in [1, n]$, and $\Lambda_i = 1 - \frac{\gamma}{s_i}$ for the other entries.

In the next section we show, by calculating the Hessian spectrum, that the manifold ($\mathcal{M}$) is a stable solution, and the zero and the mixed solutions are unstable critical points. We will also use the results of the Hessian in other parts of the paper.

A.2. Hessian eigenspace

To find the eigenvectors of the Hessian, we assume we are a small deviation away from a critical point $\tilde{\boldsymbol{\theta}} = (\tilde{U}, \tilde{W})$ along the vector $\boldsymbol{n}$. The eigenvalue equation can be written as:

$$\boldsymbol{n} = (\boldsymbol{N_U}, \boldsymbol{N_W}): \qquad \boldsymbol{H}\boldsymbol{n} = \lambda\boldsymbol{n} \equiv \begin{cases} \nabla_W L|_{\tilde{\boldsymbol{\theta}}+\boldsymbol{n}} = \lambda\boldsymbol{N_W} \\ \nabla_U L|_{\tilde{\boldsymbol{\theta}}+\boldsymbol{n}} = \lambda\boldsymbol{N_U} \end{cases} \tag{35}$$

After replacing $\boldsymbol{n}$ in the gradient equations (Eq.31), and ignoring second and higher order terms, we have the following matrix equations for $\boldsymbol{N_W}$, $\boldsymbol{N_U}$ and λ:

$$\begin{cases} (\tilde{W}\tilde{U} - \boldsymbol{I}_n)\boldsymbol{\Sigma_x}\boldsymbol{N_U^T} + (\boldsymbol{N_W}\tilde{U} + \tilde{W}\boldsymbol{N_U})\boldsymbol{\Sigma_x}\tilde{U}^T + \gamma\boldsymbol{N_W} = \lambda\boldsymbol{N_W} \\ \boldsymbol{N_W^T}(\tilde{W}\tilde{U} - \boldsymbol{I}_n)\boldsymbol{\Sigma_x} + \tilde{W}^T(\boldsymbol{N_W}\tilde{U} + \tilde{W}\boldsymbol{N_U})\boldsymbol{\Sigma_x} + \gamma\boldsymbol{N_U} = \lambda\boldsymbol{N_U} \end{cases} \tag{36}$$

To find the eigenvectors and the eigenvalues, we won't need to calculate the Hessian matrix explicitly. However, for completeness we mention here. Vectorizing the above equations lead to:

$$[(\tilde{U}\boldsymbol{\Sigma_x}\tilde{U}^T + \gamma\boldsymbol{I}_p) \otimes \boldsymbol{I}_n]\text{vec}(\boldsymbol{N_W}) + [(\boldsymbol{I}_p \otimes (\tilde{W}\tilde{U} - \boldsymbol{I}_n)\boldsymbol{\Sigma_x})\boldsymbol{K}^{(p,n)} + (\tilde{U}\boldsymbol{\Sigma_x} \otimes \tilde{W})]\text{vec}(\boldsymbol{N_U}) = \lambda\text{vec}(\boldsymbol{N_W}) \tag{37}$$

$$[(\boldsymbol{\Sigma_x}(\tilde{W}\tilde{U} - \boldsymbol{I}_n)^T \otimes \boldsymbol{I}_p)\boldsymbol{K}^{(n,p)} + (\boldsymbol{\Sigma_x}\tilde{U}^T \otimes \tilde{W}^T)]\text{vec}(\boldsymbol{N_W}) + [(\boldsymbol{\Sigma_x} \otimes \tilde{W}^T\tilde{W}) + \gamma(\boldsymbol{I}_n \otimes \boldsymbol{I}_p)]\text{vec}(\boldsymbol{N_U}) = \lambda\text{vec}(\boldsymbol{N_U})$$

where the "vec" operator reshapes a matrix to a vector, and $\otimes$ denotes the Kronecker product between two matrices. If we consider the parameter space as $\boldsymbol{\theta} \equiv \text{vec}(\boldsymbol{U}) \oplus \text{vec}(\boldsymbol{W})$, the $_{2np\times2np}$ Hessian matrix can be formed as below:

$$\boldsymbol{H} = \left(\begin{array}{c|c} \boldsymbol{H_{UU}} & \boldsymbol{H_{UW}} \\ \hline \boldsymbol{H_{WU}} & \boldsymbol{H_{WW}} \end{array}\right) \tag{38}$$

$$\boldsymbol{H_{WW}} = (\tilde{U}\boldsymbol{\Sigma_x}\tilde{U}^T + \gamma\boldsymbol{I}_p) \otimes \boldsymbol{I}_n$$
$$\boldsymbol{H_{UU}} = (\boldsymbol{\Sigma_x} \otimes \tilde{W}^T\tilde{W}) + \gamma(\boldsymbol{I}_n \otimes \boldsymbol{I}_p)$$
$$\boldsymbol{H_{WU}} = \boldsymbol{H_{UW}^T} = (\boldsymbol{I}_p \otimes (\tilde{W}\tilde{U} - \boldsymbol{I}_n)\boldsymbol{\Sigma_x})\boldsymbol{K}^{(p,n)} + (\tilde{U}\boldsymbol{\Sigma_x} \otimes \tilde{W})$$

where $\boldsymbol{K}$ is the Commutation matrix.

A.2.1. DERIVATION OF THE HESSIAN EIGENSPACE FOR THE MANIFOLD

Here we find the Hessian eigenspace for the manifold of Theorem 2.1. Replacing the equations of the manifold ($\tilde{W}\tilde{U} = \boldsymbol{I} - \gamma\boldsymbol{\Sigma_x}^{-1}$ and $\tilde{U} = \tilde{W}^T$) in Eq.36 results in the following equations for $\boldsymbol{N_W}$, $\boldsymbol{N_U}$ and λ:

$$\begin{cases} -\gamma\boldsymbol{N_U^T} + (\boldsymbol{N_W}\tilde{W}^T + \tilde{W}\boldsymbol{N_U})\boldsymbol{\Sigma_x}\tilde{W} + (\gamma - \lambda)\boldsymbol{N_W} = 0 \\ -\gamma\boldsymbol{N_W^T} + \tilde{W}^T(\boldsymbol{N_W}\tilde{W}^T + \tilde{W}\boldsymbol{N_U})\boldsymbol{\Sigma_x} + (\gamma - \lambda)\boldsymbol{N_U} = 0 \end{cases} \tag{39}$$

Since the Hessian is a real and symmetric matrix, the set of its eigenvectors are orthogonal and span $\mathbb{R}^{2np}$. A subspace of this eigenspace corresponds to $\lambda = 0$, which is the tangent space of the manifold. We will study the tangent space in more detail in the next section, but here we note that its dimension is $np - \frac{n(n+1)}{2}$, which results from the degrees of freedom (DOF) in $\tilde{W}$ in the equation of the manifold ($\tilde{W}\tilde{W}^T = $ fixed). In this section, we will find the eigenspace corresponding to $\lambda \neq 0$ (i.e. the normal space), which, as a result, has the dimension $np + \frac{n(n+1)}{2}$.

We propose two sets of solutions for the equations above for $\lambda \neq 0$. The first one satisfies:

$$\text{Set 1:} \quad \boldsymbol{N_U} = -\boldsymbol{N_W^T}, \quad \boldsymbol{N_W} = \tilde{W}\tilde{W}^T\boldsymbol{S}\tilde{W} + \boldsymbol{K}\tilde{W}_\perp, \quad \text{with } \lambda = 2\gamma. \tag{40}$$

where $\boldsymbol{S}$ is an arbitrary $n \times n$ symmetric matrix, $\boldsymbol{K}$ is an arbitrary $n \times (p - n)$ matrix, and $\tilde{W}_\perp$ is a full-rank $(p - n) \times p$ matrix whose rows are orthogonal to rows of $\tilde{W}$. (it is straightforward to check that the above satisfies Eq.39 as the term in the parentheses vanish i.e. $\boldsymbol{N_W}\tilde{W}^T - \tilde{W}\boldsymbol{N_W^T} = 0$). The total DOF in this solution is $np - \frac{n(n-1)}{2}$ resulting from DOF in $\boldsymbol{S}$ and $\boldsymbol{K}$. The second solution is:

$$\text{Set 2:} \quad \boldsymbol{N_W} = \boldsymbol{Z}\tilde{W}, \quad \boldsymbol{N_U} = \tilde{W}^T\boldsymbol{Z}, \quad \text{where } \boldsymbol{Z} \text{ satisfies } -\gamma\boldsymbol{Z}^T + 2\boldsymbol{Z}\boldsymbol{\Sigma_x} - \gamma\boldsymbol{\Sigma_x}^{-1}\boldsymbol{Z}\boldsymbol{\Sigma_x} - \lambda\boldsymbol{Z} = 0. \tag{41}$$

To solve the equation for $\boldsymbol{Z} \in \mathbb{R}^{n \times n}$ and λ, let's assume $(\boldsymbol{v}_i, s_i)$ are the eigenvectors/eigenvalues of $\boldsymbol{\Sigma_x}$ arranged in descending order, i.e. $s_1 \geq .. \geq s_n$. By replacement, we can show that n^2 solutions exists for $(\boldsymbol{Z}, \lambda)$ such that each solution corresponds to a pair of input eigenvectors $(\boldsymbol{v}_i, \boldsymbol{v}_j)$. Specifically, the solutions consists of:

$$
\begin{cases}
\boldsymbol{Z}_{ii} = C_{ii}\boldsymbol{v}_i\boldsymbol{v}_i^T, & \lambda_{ii} = 2(s_i - \gamma) & i \in [1,n] \\
\boldsymbol{Z}_{ij} = C_{ij}(\kappa_{ij}\boldsymbol{v}_i\boldsymbol{v}_j^T + \boldsymbol{v}_j\boldsymbol{v}_i^T), & \lambda_{ij} = 2s_i - \gamma(\frac{s_i}{s_j} + \kappa_{ij}) & i,j \in [1,n], i \neq j
\end{cases}
\tag{42}
$$

$$
\text{for } \kappa_{ij} = \text{sgn}(i-j)(\sqrt{1+b^2} - b), \text{ where } b = (\frac{1}{\gamma} - \frac{s_i + s_j}{2s_i s_j})|s_i - s_j|.
$$

The normalization constants C_{ij} can be found by imposing $\boldsymbol{n}_k^T \boldsymbol{n}_k = 1$, which leads to $C_{ij} = [(1 + \kappa_{ij}^2)(\omega_i + \omega_j)]^{-\frac{1}{2}}$, where $\omega_i := 1 - \frac{\gamma}{s_i}$. It is also easy to check that $\boldsymbol{n}_k^T \boldsymbol{n}_l = 0$ for $k \neq l$. Since the number of solutions in this set is n^2, we see that the total number of solutions in the two sets for the Hessian eigenspace (1 and 2) add up to the dimension of the normal space mentioned above. This shows that these two sets are the complete solutions.

Claim: For $i < j$, $\lambda_{ij} > \lambda_{ji} > 0$. (_Proof._ The left inequality results directly from writing the eigenvalues as a function of s_i and s_j noting that $s_i \geq s_j > \gamma$. For the second inequality, first note that $0 < \kappa_{ji} \leq 1$. Additionally, we have $s_j(2 - \frac{\gamma}{s_i}) > \gamma$, which proves $\lambda_{ji} = 2s_j - \gamma(\frac{s_j}{s_i} + \kappa_{ji}) > 0 \square$). _As a result, all the Hessian eigenvalues for the manifold of solution are non-negative, which makes the manifold $\mathcal{M}$ a stable solution._

A.2.2. INSTABILITY OF THE OTHER CRITICAL POINTS

Here we show that the alternate critical points found in Section A.1 are unstable by showing that their Hessian spectrum contains negative eigenvalues.

Zero solution: In this case all Λ_i are zero and we have $\tilde{W} = 0$ and $\tilde{U} = 0$. By replacement in Eq.36 we get:

$$
\begin{cases}
\boldsymbol{\Sigma_x} \boldsymbol{N}_U^T = (\gamma - \lambda)\boldsymbol{N}_W \\
\boldsymbol{N}_W^T \boldsymbol{\Sigma_x} = (\gamma - \lambda)\boldsymbol{N}_U
\end{cases}
\tag{43}
$$

Since $\boldsymbol{\Sigma_x}$ is full-rank, to have non-zero solutions to the above we should have $\lambda \neq \gamma$. Replacing $\boldsymbol{N}_U$ from the second line in the first one leads to the equation $(\boldsymbol{\Sigma_x}^2 - (\gamma - \lambda)^2 \boldsymbol{I}_n)\boldsymbol{N}_W = 0$. For a non-zero solution to exist, the determinant of the term in the parentheses should be zero. This is equivalent to the following characteristic equations: $s_i^2 - (\gamma - \lambda)^2 = 0$, for $i \in [1,n]$. Recalling from the assumptions that $s_i > \gamma$, there is a positive and a negative λ associated to each s_i: $\lambda_+ = \gamma + s_i$, and $\lambda_- = \gamma - s_i$. _Therefore, the zero solution is an unstable saddle point._

Mixed solutions: Recall from Section A.1 that for the mixed solution $\boldsymbol{W} = \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{Q}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with some entries equal to zero. For a given mixed solution, let $\mathcal{I} \subset [1,n]$ denote the indices of the zero entries of $\boldsymbol{\Lambda}$, and $\mathcal{J} = \mathcal{I}^\perp$ the indices of the non-zero diagonal entries. We can show that for the mixed solution, a part of the Hessian eigenspace (corresponding to indices $\mathcal{J}$) are similar to that of the manifold described in Section A.2.1. For example, it is easy to check that for any $j \in \mathcal{J}$ and $\boldsymbol{Z}_{jj} = C_{jj}\boldsymbol{v}_j\boldsymbol{v}_j^T$, the solution $\boldsymbol{N}_W = \boldsymbol{Z}_{jj}\tilde{W}$, $\boldsymbol{N}_U = \tilde{W}^T\boldsymbol{Z}_{jj}$ satisfy the Hessian equations (Eq.36) with positive eigenvalues $\lambda_{jj} = 2(s_j - \gamma)$. Similar to the case of the manifold, this part of the eigenspace corresponds to positive eigenvalues. However, in the mixed case, other solutions to the Hessian equation exist that have negative eigenvalues. Specifically, if $\boldsymbol{q}_i$ are columns of $\boldsymbol{Q}$, for any $i \in \mathcal{I}$ and $\boldsymbol{\alpha} \in \text{span}(\{\boldsymbol{q}_r\})$ for $r \in [1,p] - \mathcal{J}$, the solution $\boldsymbol{N}_W = \boldsymbol{N}_U^T = \boldsymbol{v}_i\boldsymbol{\alpha}^T$ satisfy Eq.36 with the eigenvalue $\lambda = \gamma - s_i$, which is negative. Hence, in this case we have both negative and positive eigenvalues. _This proves that the mixed solutions are unstable saddle points._

A.3. Summary of the proof of Theorem 2.1

In Section A.1 we found all the critical points and classified them into three sets consisting of the manifold ($\mathcal{M}$), the zero, and the mixed solutions. In Section A.2.1, we showed that the Hessian eigenvalues are non-negative for the manifold (see the Claim at the end of the section). Finally, in Section A.2.2, we showed that the other critical points are unstable saddle points. This proves that $\mathcal{M}$ consists of all the stable critical points in the parameter space. $\square$

B. Tangent Space to the Manifold

In this section, we describe the first order differential geometry of the manifold ($\mathcal{M}$) which will allow us to study the dynamics near the manifold. In the two equations describing the manifold in Theorem 2.1, $\tilde{U}$ is simply the transpose of $\tilde{W}$, and hence the manifold can be studied by considering matrix $\tilde{W} \in \mathbb{R}^{n \times p}$ in the equation $\tilde{W}\tilde{W}^T = I_n - \gamma \Sigma_x^{-1}$. This defines a matrix manifold in $\mathbb{R}^{n \times p}$. In the case of $\gamma = 0$, this simplifies to the Steifel manifold for which there are known results (see Edelman et al. (1998) and Absil et al. (2009)). Inspired by those results, here we consider a more general but similar case of $\gamma \neq 0$. The next lemma describes the tangent space to the manifold.

Lemma 2.2. (Tangent space) *The local tangent space to the manifold at point $(\tilde{W}^T, \tilde{W})$ is spanned by vectors $t = (T_W^T, T_W)$ with $T_W = \tilde{W}\tilde{W}^T \Omega \tilde{W} + K \tilde{W}_\perp$, where Ω is an arbitrary $n \times n$ skew-symmetric matrix, K is an arbitrary $n \times (p-n)$ matrix, and $\tilde{W}_\perp$ is a full-rank $(p-n) \times p$ matrix whose rows are orthogonal to rows of $\tilde{W}$.*

Proof. First, we show that any vector t of the form above belongs to the tangent space of the manifold. This is equivalent to showing that if we are on the manifold and get displaced by vector $\varepsilon t = (\varepsilon T_W^T, \varepsilon T_W)$ for a small ε, we stay on the manifold to the first order of ε (i.e. the equations that define the manifold still hold to that order). From the two equations, the transpose equation ($\tilde{U} = \tilde{W}^T$) is automatically satisfied as we have $T_U = T_W^T$. For the other equation ($\tilde{W}\tilde{W}^T = I - \gamma \Sigma_x^{-1}$), we show that the change in WW^T is of the second order:

$$\begin{aligned}
\Delta(WW^T) &= \varepsilon \tilde{W} T_W^T + \varepsilon T_W \tilde{W}^T + \mathcal{O}(\varepsilon^2) \\
&= \varepsilon \tilde{W}\tilde{W}^T \Omega^T \tilde{W}\tilde{W}^T + \varepsilon \tilde{W}\tilde{W}_\perp^T K^T + \varepsilon \tilde{W}\tilde{W}^T \Omega \tilde{W}\tilde{W}^T + \varepsilon K \tilde{W}_\perp \tilde{W}^T + \mathcal{O}(\varepsilon^2) \\
&= \mathcal{O}(\varepsilon^2)
\end{aligned} \tag{44}$$

where in going from the second to the third line we used $\Omega = -\Omega^T$ and $\tilde{W}_\perp \tilde{W}^T = 0$.

Next, we demonstrate that the dimension of the vector space spanned by t is the same as the dimension of the tangent space. The dimension of the tangent space can be inferred from the difference between the parameters and the constraints in $\tilde{W}\tilde{W}^T = I - \gamma \Sigma_x^{-1}$, which are np and $n(n+1)/2$ respectively. Hence, the dimension of the tangent space is $np - n(n+1)/2$. To find the dimension of the space spanned by t, note that the transformation from (Ω, K) to T_W is one-to-one. This is simply because each of the mappings $\Omega \to \tilde{W}\tilde{W}^T \Omega \tilde{W}$ and $K \to K \tilde{W}_\perp$ are one-to-one (as $\tilde{W}$ and $\tilde{W}_\perp$ are full rank), and their ranges do not overlap (as the rows of $\tilde{W}$ and $\tilde{W}_\perp$ are orthogonal). As a result, the dimension of the space spanned by T_W is equal to the dimension of the space spanned by the skew-symmetric Ω, which is $n(n-1)/2$, added to that of K, which is $n(p-n)$. This becomes the same as the dimension of tangent space mentioned above. Since we already showed that any vector t lies in the tangent space, the set of t's span the tangent space. This completes the proof. $\square$

With the definition of the Euclidean inner product in Eq.2, we can define notions of orthogonality and projection. In the case of the two layer network ($L = 2$) with weights U and W, the inner product for two vectors $e = (E_U, E_W)$ and $f = (F_U, F_W)$ becomes:

$$e^T f = \mathrm{tr}(E_W^T F_W) + \mathrm{tr}(E_U^T F_U) \tag{45}$$

The normal space can be defined as the space orthogonal to the tangent space. This is the same as the eigenspace of the Hessian corresponding to positive eigenvalues and was described in Section A.2.1 (similarly, vectors in the tangent space corresponds to Hessian eigenvectors with zero eigenvalue, i.e. $Ht = 0$). Finally, the inner product of an arbitrary vector g with vector t in the tangent space can be found from:

$$t^T g = \mathrm{tr}(T_W^T G_W) + \mathrm{tr}(T_U^T G_U) = \mathrm{tr}(T_W(G_W^T + G_U)) \tag{46}$$

where T_W takes an appropriate form from Lemma 2.2. A natural basis for the tangent space can be constructed from matrices $\Omega_{rs} := (v_r v_s^T - v_s v_r^T)/\sqrt{2\omega_r \omega_s (\omega_r + \omega_s)}$, for $r, s \in [1, n]$ and $r > s$, as well as matrices K_{uv} with components $[K_{uv}]_{ij} = \delta_{ij}^{uv}$ ($u, i \in [1, n]$, $v, j \in [1, p-n]$). Recall that $\omega_i = 1 - \gamma/s_i$ and (v_i, s_i) are eigenvector/eigenvalue pairs of the input covariance, and the normalization constants ensure $t^T t = 1$. In Section D, we use these results to project the gradient onto the tangent space.

C. Calculation of the Fluctuations

In this section, we present additional results and derivations for the fluctuations in the normal space.

C.1. Derivation of the SDE for normal deviations

Here, we will derive the stochastic differential equation (SDE) for $\boldsymbol{\theta}_N$ in Eq.7. Without loss of generality, at a given time, we take $\tilde{\boldsymbol{\theta}}$ to be the closest point to $\boldsymbol{\theta}$, and also put the origin on it. Projecting both sides of the SGD equation (Eq.4) to the normal space of the manifold, and by defining $\boldsymbol{g}_N(\boldsymbol{x}; \boldsymbol{\theta}_N) = \Pi_N(\boldsymbol{g}(\boldsymbol{x}; \boldsymbol{\theta}_N))$ to be the normal projection of the gradient, we have:

$$\Delta\boldsymbol{\theta}_N = -\eta\boldsymbol{g}_N = -\eta\langle\boldsymbol{g}\rangle_x - \eta\boldsymbol{e}$$

where $\boldsymbol{e} = \boldsymbol{g}_N - \langle\boldsymbol{g}\rangle_x$ is a random variable with zero mean, and $\langle\boldsymbol{g}_N\rangle_x = \langle\boldsymbol{g}\rangle_x$ is the average gradient is in the normal direction. The above is the update in one training step which we take to be equivalent to the continuous time-difference $\Delta t_{step} \equiv \eta$. Over large timescales ($\Delta t = n\eta$, with $n \gg 1$), the update equation becomes:

$$\Delta\boldsymbol{\theta}_N = -n\eta\langle\boldsymbol{g}\rangle_x - \boldsymbol{e}' \quad \text{(n steps)} \tag{47}$$

Here $\boldsymbol{e}' = \eta\sum_{i=1}^n \boldsymbol{e}_i$ is a random variable with $\langle\boldsymbol{e}'\rangle = 0$ and $\text{var}(\boldsymbol{e}') = n\eta^2\text{cov}(\boldsymbol{e}) = \Delta t\,\eta\,\text{cov}(\boldsymbol{e})$. As a result, we can approximate it as $\boldsymbol{e}' \sim \sqrt{\eta\,\text{cov}(\boldsymbol{e})}\boldsymbol{B}_{\Delta t}$ where $\boldsymbol{B}_t$ is a multi-dimensional Brownian motion (Wiener process) with properties such as independent Gaussian increments and a variance proportional to the time difference, i.e. $\boldsymbol{B}_{t+u} - \boldsymbol{B}_t \sim \mathcal{N}(\boldsymbol{0}, u\boldsymbol{I}_{2np})$. In the limit of $\Delta t \to 0$, we have the following continuous stochastic differential equation (SDE):

$$d\boldsymbol{\theta}_N = -\langle\boldsymbol{g}\rangle_x dt + \sqrt{\eta}\boldsymbol{C}d\boldsymbol{B}_t \quad \text{(continuous)} \tag{48}$$

where $\boldsymbol{C} \equiv \sqrt{\text{cov}(\boldsymbol{e})}$. Since $\boldsymbol{e}$ in general is a function of $\boldsymbol{\theta}_N$, we can write $\boldsymbol{C}(\boldsymbol{\theta}_N) = \boldsymbol{C}(0) + \mathcal{O}(\boldsymbol{\theta}_N)$. Near the manifold we also have $\langle\boldsymbol{g}\rangle_x = \boldsymbol{H}\boldsymbol{\theta}_N + \mathcal{O}(\theta_N^2)$. As a first approximation, we replace $\boldsymbol{C}(\boldsymbol{\theta}_N)$ by $\boldsymbol{C}(0)$, which means the driver of the stochasticity in the equation becomes the gradient on the manifold, i.e. $\tilde{\boldsymbol{g}}(\boldsymbol{x}) := \boldsymbol{g}_N(\boldsymbol{x}; \boldsymbol{0})$. Note that as the next section shows, under the stationary solution $\theta_N \sim \sqrt{\eta}$, which justifies the approximation in the second term of the equation for small η (see also Mandt et al. (2017)).

C.2. Solution of the SDE for normal deviations

The solution to the SDE in Eq.7 (aka Ornstein Uhlenbeck process) has been discussed previously (Gardiner et al., 1985). For completeness, we provide it here. The process was defined by:

$$d\boldsymbol{\theta}_N(t) = -\boldsymbol{H}\boldsymbol{\theta}_N(t)dt + \sqrt{\eta}\boldsymbol{C}d\boldsymbol{B}_t \tag{49}$$

We first perform the following change of variable:

$$\boldsymbol{\theta}'_N(t) := e^{\boldsymbol{H}t}\boldsymbol{\theta}_N(t) \to \boldsymbol{\theta}_N(t) = e^{-\boldsymbol{H}t}\boldsymbol{\theta}'_N(t). \tag{50}$$

By taking the differential of the above and replacement from Eq.49 we get:

$$d\boldsymbol{\theta}'_N(t) = \boldsymbol{H}e^{\boldsymbol{H}t}\boldsymbol{\theta}_N(t)dt + e^{\boldsymbol{H}t}(-\boldsymbol{H}\boldsymbol{\theta}_N(t)dt + \sqrt{\eta}\boldsymbol{C}d\boldsymbol{B}_t) = \sqrt{\eta}e^{\boldsymbol{H}t}\boldsymbol{C}d\boldsymbol{B}_t \tag{51}$$

After integration (Ito integral) and substitution from 50, we have:

$$\boldsymbol{\theta}'_N(t) = \boldsymbol{\theta}'_N(0) + \sqrt{\eta}\int_0^t e^{\boldsymbol{H}u}\boldsymbol{C}d\boldsymbol{B}_u \to \boldsymbol{\theta}_N(t) = e^{-\boldsymbol{H}t}\boldsymbol{\theta}_N(0) + \sqrt{\eta}\int_0^t e^{-\boldsymbol{H}(t-u)}\boldsymbol{C}d\boldsymbol{B}_u \tag{52}$$

For a given initial condition, the mean is:

$$\langle\boldsymbol{\theta}_N(t)\rangle = e^{-\boldsymbol{H}t}\boldsymbol{\theta}_N(0) \tag{53}$$

which shows the stationary mean ($t \to \infty$) is zero. Correspondingly, the correlation function is:

$$\text{corr}(\boldsymbol{\theta}_N(t), \boldsymbol{\theta}_N(s)) = \langle[\boldsymbol{\theta}_N(t) - \langle\boldsymbol{\theta}_N(t)\rangle][\boldsymbol{\theta}_N(s) - \langle\boldsymbol{\theta}_N(s)\rangle]^T\rangle \tag{54}$$

$$= \eta\langle\int_0^t e^{-\boldsymbol{H}(t-u)}\boldsymbol{C}d\boldsymbol{B}_u[\int_0^s e^{-\boldsymbol{H}(s-v)}\boldsymbol{C}d\boldsymbol{B}_v]^T\rangle$$

$$= \eta\int_0^{\min(t,s)} e^{-\boldsymbol{H}(t-u)}\boldsymbol{C}\boldsymbol{C}^T e^{-\boldsymbol{H}^T(s-u)}du$$

where in the last line we used a property of Ito integral. Now recall the change of variable from the main text $\boldsymbol{\rho} = \boldsymbol{N}^T \boldsymbol{\theta}_N$ (Eq.9), which means $\text{corr}(\boldsymbol{\rho}(t), \boldsymbol{\rho}(s)) = \boldsymbol{N}^T \text{corr}(\boldsymbol{\theta}_N(t), \boldsymbol{\theta}_N(s))\boldsymbol{N}$. Additionally, we have the decomposition $\boldsymbol{H} = \boldsymbol{N}\boldsymbol{\lambda}\boldsymbol{N}^T$, where $\boldsymbol{\lambda} = \text{diag}(\{\lambda_k\})$. Replacing these and $\boldsymbol{C}\boldsymbol{C}^T = \langle \tilde{g}\tilde{g}^T \rangle_x$ in the above, the integral can be calculated per component. Further, taking $t, s \to \infty$ yields the correlation function for the stationary solution:

$$\langle \rho_k(t)\rho_l(s) \rangle = \frac{\eta \langle \boldsymbol{n}_k^T \tilde{\boldsymbol{g}} \, \boldsymbol{n}_l^T \tilde{\boldsymbol{g}} \rangle_x}{\lambda_k + \lambda_l} e^{-\lambda_k(t-s)}, \ (t \geqslant s). \tag{55}$$

The covariance equation in the main text (Eq.10) results from the above under $t = s$.

C.3. Projection of the gradient onto the normal space (Proof of Proposition 2.3)

In Section A.2.1 we found the normal space which consisted of two sets of solutions (Eqns. 40 and 41). Here we show that $\tilde{g}(\boldsymbol{x})$ has no projection the first set. Recall that for the first solution we had: $\boldsymbol{n} = (-\boldsymbol{N}_W^T, \boldsymbol{N}_W)$ where $\boldsymbol{N}_W \tilde{\boldsymbol{W}}^T = \tilde{\boldsymbol{W}} \boldsymbol{N}_W^T$. Using $\tilde{g}(\boldsymbol{x}) = (\tilde{\boldsymbol{W}}^T \boldsymbol{Z}_x, \boldsymbol{Z}_x \tilde{\boldsymbol{W}})$ from Eq.8, and the definition of the inner product (Eq.45), we have:

$$\text{Set 1:} \quad \boldsymbol{n}^T \tilde{g}(\boldsymbol{x}) = \text{tr}\left(\boldsymbol{N}_W^T \boldsymbol{Z}_x \tilde{\boldsymbol{W}}\right) - \text{tr}\left(\boldsymbol{N}_W \tilde{\boldsymbol{W}}^T \boldsymbol{Z}_x\right) = \text{tr}\left(\boldsymbol{N}_W^T \boldsymbol{Z}_x \tilde{\boldsymbol{W}}\right) - \text{tr}\left(\tilde{\boldsymbol{W}} \boldsymbol{N}_W^T \boldsymbol{Z}_x\right) = 0$$

where in the last equality we used the permutation property of the trace. This leaves us with the second set of solutions which consisted of $\boldsymbol{n}_{ij} = (\tilde{\boldsymbol{W}}^T \boldsymbol{Z}_{ij}, \boldsymbol{Z}_{ij}\tilde{\boldsymbol{W}})$ for $i, j \in [1, n]$ (Eq.42). The projection on these eigenvectors can be found similarly using the definition of inner product. By defining the coefficients

$$S_{rs}^{ij} := \frac{\kappa_{ij}}{s_r} + \frac{1}{s_s} \text{ (noting } \kappa_{ii} \equiv 0), \quad \text{and} \quad \omega_i := 1 - \frac{\gamma}{s_i}, \tag{56}$$

(see Section A.2.1 for definitions of κ_{ij} and C_{ij}), we will have:

$$\text{Set 2:} \quad \boldsymbol{n}_{ii}^T \tilde{\boldsymbol{g}} = \gamma \sqrt{2\omega_i}(1 - \frac{x_i^2}{s_i}), \quad \boldsymbol{n}_{ij}^T \tilde{\boldsymbol{g}}_{(i \neq j)} = -\gamma(\omega_i + \omega_j)C_{ij} S_{ij}^{ij} x_i x_j, \tag{57}$$

where $x_i := \boldsymbol{v}_i^T \boldsymbol{x}$. Hence the projections on set 2 are in general non-zero. This completes the proof.

C.4. Components of the fluctuation matrix

By replacing the projections from Eq.57 in Eq.10, the components of the covariance matrix can be calculated as:

$$\langle \rho_{ii}\rho_{pp} \rangle = \frac{\eta \gamma^2 \sqrt{\omega_i \omega_p}}{s_i + s_p - 2\gamma}\left(\frac{\langle x_i^2 x_p^2 \rangle_x}{s_i s_p} - 1\right), \qquad \langle \rho_{ii}\rho_{pq} \rangle_{(p \neq q)} = \langle \rho_{ij}\rho_{pq} \rangle_{(i \neq j)} = 0 \tag{58}$$

$$\langle \rho_{ij}\rho_{pq} \rangle_{(i \neq j, p \neq q)} = \frac{\eta \gamma^2 (\omega_i + \omega_j)(\omega_p + \omega_q)}{\lambda_{ij} + \lambda_{pq}} C_{ij} C_{pq} S_{ij}^{ij} S_{pq}^{pq} \langle x_i x_j x_p x_q \rangle_x$$

where $i, j, p, q \in [1, n]$.

C.5. Fluctuation of the representation norm

Consider $\boldsymbol{h}_s = \tilde{\boldsymbol{U}}\boldsymbol{v}_s$ to be the representation of stimulus $\boldsymbol{v}_s$. According to Proposition 2.3, displacement from the manifold along vector $\boldsymbol{\rho} = \sum \rho_k \boldsymbol{n}_k$ in the normal space changes the first layer weight by $\Delta \boldsymbol{U} = \tilde{\boldsymbol{W}}^T \sum_{i,j} \rho_{ij} C_{ij}(\kappa_{ij}\boldsymbol{v}_i\boldsymbol{v}_j^T + \boldsymbol{v}_j\boldsymbol{v}_i^T)$. This correspondingly changes the representation by $\Delta \boldsymbol{h}_s = \Delta \boldsymbol{U}\boldsymbol{v}_s$. The change in the representation norm can therefore be calculated as:

$$\hat{\boldsymbol{h}}_s^T \Delta \boldsymbol{h}_s = \sqrt{\omega_s}\boldsymbol{v}_s^T \sum_{i,j=1}^{n} \rho_{ij} C_{ij}(\kappa_{ij}\boldsymbol{v}_i\boldsymbol{v}_j^T + \boldsymbol{v}_j\boldsymbol{v}_i^T)\boldsymbol{v}_s = \frac{\rho_{ss}}{\sqrt{2}} \tag{59}$$

where we used $\tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^T \boldsymbol{v}_s = \omega_s \boldsymbol{v}_s$ and $C_{ss} = 1/\sqrt{2\omega_s}$. The variance in the norm becomes:

$$\text{var}(|\boldsymbol{h}_s|) = \frac{\langle \rho_{ss}^2 \rangle}{2} \tag{60}$$

which is the first part of Eq.12 in the main paper. The right hand side of Eq.12 results simply by taking $i = p = s$ in the first equation of Eq.58 to calculate $\langle \rho_{ss}^2 \rangle$.

D. Tangential Projection of the Gradient (Proof of Theorem 2.5)

We first approximate the gradient near the manifold to the first order of ρ. We assume we are at point $\tilde{\theta} + \rho_k n_k$, where $\tilde{\theta}$ lies on the manifold, and $n_k = (\tilde{W}^T Z_k, Z_k \tilde{W})$ is a Hessian eigenvector (Eq.11). The gradient at this point can be calculated by replacing $W = (I + \rho_k Z_k)\tilde{W}$ and $U = \tilde{W}^T(I + \rho_k Z_k)$ in Eq.31. For $\rho_k \ll 1$, we will ignore second and higher order terms, to have:

$$g(x; \tilde{\theta} + \rho_k n_k) = (G_U, G_W); \tag{61}$$

$$
\begin{aligned}
G_W &= (WU - I)xx^T U^T + \gamma W \\
&= [(I + \rho_k Z_k)\tilde{W}\tilde{W}^T(I + \rho_k Z_k) - I]xx^T(I + \rho_k Z_k^T)\tilde{W} + \gamma(I + \rho_k Z_k)\tilde{W} + \mathcal{O}(\rho_k^2) \\
&= \gamma(I - \Sigma_x^{-1}xx^T)\tilde{W} + \rho_k[(Z_k\tilde{W}\tilde{W}^T + \tilde{W}\tilde{W}^T Z_k)xx^T - \gamma\Sigma_x^{-1}xx^T Z_k^T + \gamma Z_k]\tilde{W} + \mathcal{O}(\rho_k^2)
\end{aligned}
$$

$$
\begin{aligned}
G_U &= W^T(WU - I)xx^T + \gamma U \\
&= \tilde{W}^T(I + \rho_k Z_k^T)[(I + \rho_k Z_k)\tilde{W}\tilde{W}^T(I + \rho_k Z_k) - I]xx^T + \gamma\tilde{W}^T(I + \rho_k Z_k) + \mathcal{O}(\rho_k^2) \\
&= \gamma\tilde{W}^T(I - \Sigma_x^{-1}xx^T) + \rho_k\tilde{W}^T[(Z_k\tilde{W}\tilde{W}^T + \tilde{W}\tilde{W}^T Z_k)xx^T - \gamma Z_k^T\Sigma_x^{-1}xx^T + \gamma Z_k] + \mathcal{O}(\rho_k^2)
\end{aligned}
$$

As was shown Section B, the dependency of the projection on the gradient is via $G_W^T + G_U$, which can be calculated as:

$$
\begin{aligned}
G_W^T + G_U &= -\gamma\rho_{ij}\tilde{W}^T(Z_{ij}xx^T\Sigma_x^{-1} + Z_{ij}^T\Sigma_x^{-1}xx^T) + 2\tilde{W}^T\text{Sym}(A) + \mathcal{O}(\rho_{ij}^2) \\
&= -\gamma\rho_{ij}\tilde{W}^T C_{ij} \sum_{q \in [1,n]} (S_{qj}^{ij}x_j x_q v_i v_q^T + S_{iq}^{ij}x_i x_q v_j v_q^T) + 2\tilde{W}^T\text{Sym}(A) + \mathcal{O}(\rho_{ij}^2). \tag{62}
\end{aligned}
$$

In the above, $A = \gamma(I_n - xx^T\Sigma_x^{-1}) + \rho_{ij}[(Z_{ij}\tilde{W}\tilde{W}^T + \tilde{W}\tilde{W}^T Z_{ij})xx^T + \gamma Z_{ij}]$, and S_{rs}^{ij} are coefficients defined in Eq.56. In the second line, we replaced $Z_k \equiv Z_{ij} = C_{ij}(\kappa_{ij}v_i v_j^T + v_j v_i^T)$ (Eq.11), $x = \sum_{q \in [1,n]} x_q v_q$, and used $\Sigma_x^{-1}v_q = s_q^{-1}v_q$ and $v_p^T v_q = \delta_p^q$.

The inner product of the gradient on a vector in the tangent space becomes $t^T g = \text{tr}(T_W(G_W^T + G_U))$ (Eq.46), where from Lemma 2.2 we have $T_W = \tilde{W}\tilde{W}^T\Omega\tilde{W} + K\tilde{W}_\perp$. Replacing the $G_W^T + G_U$ term from Eq.62, we first see that the gradient has no projection on the second term of the tangent vector containing K, as we have $\tilde{W}_\perp\tilde{W}^T = 0$. Further, the contribution of the term $2\tilde{W}\text{Sym}(A)$ in Eq.62 also disappears as the trace of the product of symmetric and skew-symmetric matrices is zero. This leaves us with calculating the projection on tangent vectors corresponding to Ω. As described in Section B, the associated basis vectors are $t_{rs} = (T_W^T, T_W)$, where $T_W = \tilde{W}\tilde{W}^T\Omega_{rs}\tilde{W}$, and $\Omega_{rs} := (v_r v_s^T - v_s v_r^T)/\sqrt{2\omega_r\omega_s(\omega_r + \omega_s)}$. The projection on t_{rs} becomes:

$$
\begin{aligned}
t_{rs}^T g(x; \tilde{\theta} + \rho_k n_k) &= \text{tr}(\tilde{W}\tilde{W}^T\Omega_{rs}\tilde{W}(G_W^T + G_U)) \\
&= \frac{\gamma\rho_{ij}C_{ij}\sqrt{\omega_s\omega_r}}{\sqrt{2(\omega_s + \omega_r)}}[x_j(S_{sj}^{ij}x_s\delta_i^r - S_{rj}^{ij}x_r\delta_i^s) + x_i(S_{is}^{ij}x_s\delta_j^r - S_{ir}^{ij}x_r\delta_j^s)] + \mathcal{O}(\rho_{ij}^2) \\
&= \rho_{ij}\sqrt{2(\omega_s + \omega_r)}\mathcal{G}_{ij}^{s,r}(x) + \mathcal{O}(\rho_{ij}^2) \tag{63}
\end{aligned}
$$

where in the last line we defined the rank-3 tensor $\mathcal{G}$ with components that are:

$$\mathcal{G}_{ij}^{s,r}(x) = -\mathcal{G}_{ij}^{r,s}(x) = \frac{\gamma C_{ij}\sqrt{\omega_r\omega_s}}{2(\omega_r + \omega_s)}[x_j(S_{sj}^{ij}x_s\delta_i^r - S_{rj}^{ij}x_r\delta_i^s) + x_i(S_{is}^{ij}x_s\delta_j^r - S_{ir}^{ij}x_r\delta_j^s)]. \tag{64}$$

(note that the lower index $[k]$ is a composite index and we use it interchangeably with ij, i.e. $\mathcal{G}_k^{s,r} \equiv \mathcal{G}_{ij}^{s,r}$). The projection of the gradient on the tangent space becomes $g_T = \sum_{r>s}(t_{rs}^T g)\, t_{rs} = (G_{U_T}, G_{U_T}^T)$, where:

$$
\begin{aligned}
G_{U_T}(x; \tilde{\theta} + \rho_{ij}n_{ij}) &= \rho_{ij}\tilde{W}^T \sum_{r>s} \mathcal{G}_{ij}^{s,r}(\sqrt{\frac{\omega_r}{\omega_s}}v_s v_r^T - \sqrt{\frac{\omega_s}{\omega_r}}v_r v_s^T) + \mathcal{O}(\rho_{ij}^2) \tag{65} \\
&= \tilde{W}^T(\tilde{W}\tilde{W}^T)^{-\frac{1}{2}}(\rho_{ij}\mathcal{G}_{ij}^{:,:})(\tilde{W}\tilde{W}^T)^{\frac{1}{2}} + \mathcal{O}(\rho_{ij}^2)
\end{aligned}
$$

(In the second line above we used $\mathcal{G}_{ij}^{:,:} = \sum_{s,r} \mathcal{G}_{ij}^{r,s}v_s v_r^T$). Due to linearity, the tangential projection at point $N\rho = \sum_k \rho_k n_k$ can be found by summing over k, leading to Eq.14 in the main text:

$$G_{U_T}(x; \tilde{\theta} + N\rho) = \sum_k G_{U_T}(x; \tilde{\theta} + \rho_k n_k) = \tilde{W}^T(\tilde{W}\tilde{W}^T)^{-\frac{1}{2}}(\sum_k \rho_k\mathcal{G}_k^{:,:})(\tilde{W}\tilde{W}^T)^{\frac{1}{2}} + \mathcal{O}(|\rho|^2) \tag{66}$$

Next, we will find the effect of the tangential gradient on the representations. For a given stimulus $\boldsymbol{s}$, the change in its representation due to the tangential gradient is $\Delta \boldsymbol{h} = -\eta \boldsymbol{G}_{\boldsymbol{U}_T} \boldsymbol{s}$. Replacing $\boldsymbol{G}_{\boldsymbol{U}_T}$ from Eq.66, and using $\boldsymbol{h} = \boldsymbol{U}\boldsymbol{s} \approx \tilde{\boldsymbol{W}}^T \boldsymbol{s}$ (for small deviations from the manifold), we have:

$$\boldsymbol{h}^T \Delta \boldsymbol{h} = -\eta \boldsymbol{s}^T (\tilde{\boldsymbol{W}} \tilde{\boldsymbol{W}}^T)^{\frac{1}{2}} (\sum_k \rho_k \boldsymbol{\mathcal{G}}_k^{:,:}) (\tilde{\boldsymbol{W}} \tilde{\boldsymbol{W}}^T)^{\frac{1}{2}} \boldsymbol{s} = 0 \tag{67}$$

The second equality results from the fact that the quadratic form of a skew-symmetric matrix ($\boldsymbol{\mathcal{G}}_k^{:,:}$) is zero. This shows that, as expected, the tangential projection of the gradient is equivalent to a small rigid-body rotation of the representations around the origin. Further, $\tilde{\boldsymbol{W}}^T$ on the left hand side of Eq.66 makes $\Delta \boldsymbol{h}$ to be a linear combination of the columns of $\tilde{\boldsymbol{W}}^T$, keeping the representations in the column-space of $\tilde{\boldsymbol{W}}^T$ (note also that the column-space of $\tilde{\boldsymbol{W}}^T$ stays fixed over a tangential update, which is again due to the $\tilde{\boldsymbol{W}}^T$ term on the left hand side of Eq.66). To characterize this rotation, we measure its effect on the n representations vectors. Specifically, we define $\Delta \varphi_{sr}$ to be the the pairwise angular displacement of representation $\boldsymbol{h}_s \approx \tilde{\boldsymbol{W}}^T \boldsymbol{v}_s$ toward $\boldsymbol{h}_r \approx \tilde{\boldsymbol{W}}^T \boldsymbol{v}_r$:

$$\Delta \varphi_{sr} := \frac{\tilde{\boldsymbol{h}}_r^T \Delta \boldsymbol{h}_s}{|\tilde{\boldsymbol{h}}_r||\tilde{\boldsymbol{h}}_s|} = \eta \sum_k \rho_k \boldsymbol{\mathcal{G}}_k^{s,r} + \mathcal{O}(|\boldsymbol{\rho}|^2), \tag{68}$$

where we replaced $\Delta \boldsymbol{h}_s$ from the above and used $|\boldsymbol{h}_s| = \sqrt{\omega_s}$, and $\boldsymbol{\mathcal{G}}_k^{r,s} = \boldsymbol{v}_r^T \boldsymbol{\mathcal{G}}_k^{:,:} \boldsymbol{v}_s$. This is Eq.16 in the main text.

E. Derivation of the Diffusion Coefficients for the Case with a Frequent Stimulus

Here we derive approximate analytical solutions for the diffusion coefficients for the case with a frequent stimulus.

E.1. Approximation of the diffusion summation

Since in the D_{sr} summation in Eq.18, $k \equiv (ij)$ and $l \equiv (pq)$ are composite indices with $i, j, p, q \in [1, n]$, the summation is in general performed over n^4 terms (first line of Eq.69 below). However, as we'll see next, most of the terms are zero and under different limits, we can further constrain it to a subset of indices.

$n \gg 1$: If we consider the equation for $\boldsymbol{\mathcal{G}}_{ij}^{r,s}$ for a given r and s (Eqn, 15), we see that it is non-zero only when at least either i or j are equal to r or s. For a fixed r and s, looping over $i, j \in [1, n]$ leads to a maximum of two non-zero components over the $\boldsymbol{\mathcal{G}}_{ii}^{r,s}$ terms, and up to $\approx 4n$ non-zero components over the $\boldsymbol{\mathcal{G}}_{ij(i \neq j)}^{r,s}$ terms. Hence, for large n, the terms corresponding to the first set can be ignored. This leaves us with a summation over indices (ij, pq) for $i \neq j$ and $p \neq q$, which consists of up to $\approx 16n^2$ non-zero terms. Additionally, both $\bar{\bar{\boldsymbol{\mathcal{G}}}}_{ij,pq}^{r,s}$ and $\langle \rho_{ij} \rho_{pq} \rangle$ are proportional to terms like $\langle x_{r/s} x_{i/j} x_{r/s} x_{p/q} \rangle_x$, which are zero if an odd number of indices are equal. This leaves us with a summation over indices of (ij, ij) and (ij, ji), where $(i, j) \in \mathbb{C}_{[ij]} \cup \mathbb{C}_{[ji]}$, for set $\mathbb{C}_{[ij]} := \{(i, j) | i \in \{r, s\}, j \in [1, n] \backslash \{r, s\}\}$ (this is shown in the second line of Eq.69 below). Since $\mathbb{C}_{[ij]}$ contains $\approx 2n$ terms, the whole summation will have up to $\approx 8n$ terms.

$\alpha \ll 1$: For small α, we can further limit the summation to indices (ij, ij) for $i > j$. This is because the first order correction to the summands in Eq.18 is: $\approx \langle \rho_k \rho_l \rangle|_{\alpha=0} \bar{\bar{\boldsymbol{\mathcal{G}}}}_{k,l}^{r,s}|_{\mathcal{O}(\alpha)} + \langle \rho_k \rho_l \rangle|_{\mathcal{O}(\alpha)} \bar{\bar{\boldsymbol{\mathcal{G}}}}_{k,l}^{r,s}|_{\alpha=0}$, and so we can limit the sum to indices for which $\langle \rho_k \rho_l \rangle|_{\alpha=0}$ or $\bar{\bar{\boldsymbol{\mathcal{G}}}}_{k,l}^{r,s}|_{\alpha=0}$ are non-zero. This corresponds to the previous Gaussian stimuli case and the terms were calculated in Section 3 of the main text. To summarize:

$$\begin{aligned} D_{sr} &= (\eta^2/2) \sum_{i,j,p,q=1}^{n} \langle \rho_{ij} \rho_{pq} \rangle \bar{\bar{\boldsymbol{\mathcal{G}}}}_{ij,pq}^{r,s} \\ &\approx (\eta^2/2) \sum_{(i,j) \in \mathbb{C}_{[ij]} \cup \mathbb{C}_{[ji]}} (\langle \rho_{ij}^2 \rangle \bar{\bar{\boldsymbol{\mathcal{G}}}}_{ij,ij}^{r,s} + \langle \rho_{ij} \rho_{ji} \rangle \bar{\bar{\boldsymbol{\mathcal{G}}}}_{ij,ji}^{r,s}) \qquad n \gg 1 \\ &\approx (\eta^2/2) \sum_{(i,j) \in (\mathbb{C}_{[ij]} \cup \mathbb{C}_{[ji]}) \wedge (i>j)} \langle \rho_{ij}^2 \rangle \bar{\bar{\boldsymbol{\mathcal{G}}}}_{ij,ij}^{r,s} \qquad \alpha \ll 1, n \gg 1 \end{aligned} \tag{69}$$

We see that for a given s and r, we will have up to $\approx 2n$ terms in the summation under the $n \gg 1$ and $\alpha \ll 1$ limit.

To fully characterize the diffusion in the representation space, we only need to find two coefficients D_{ab} ($= D_{ba}$) and D_{bc} ($= D_{cb}$), where $a = 1$, $b \neq c \in [2, n]$. Without loss of generality, we take $b = 2$ and $c = 3$. We also take $d = 4$ to be another index within the background subspace. In the next two sections, we calculate these coefficients under two limits.

E.2. Calculation for large n and small α

For small α, D_{ba} and D_{bc} can be found by replacement in the last line of Eq.69:

$$D_{ba} \approx \frac{n\eta^2}{2}(\langle\rho_{db}^2\rangle\bar{\bar{\boldsymbol{\mathcal{G}}}}_{db,db}^{a,b} + \langle\rho_{da}^2\rangle\bar{\bar{\boldsymbol{\mathcal{G}}}}_{da,da}^{a,b}) \qquad (\alpha \ll 1, n \gg 1)$$

$$D_{bc} \approx \frac{n\eta^2}{2}(\langle\rho_{db}^2\rangle\bar{\bar{\boldsymbol{\mathcal{G}}}}_{db,db}^{c,b} + \langle\rho_{dc}^2\rangle\bar{\bar{\boldsymbol{\mathcal{G}}}}_{dc,dc}^{c,b}) \qquad (\alpha \ll 1, n \gg 1)$$

The relevant components of tensor $\boldsymbol{\mathcal{G}}$ can be calculated from Eq.15 for the case of frequent stimulus. For $\gamma \ll 1$ we have $\omega_i \approx 1$, $\kappa_{ij(i>j>1)} = 1$, $C_{ij(i\neq j>1)} \approx 1/2$, $\langle x_i^2 x_j^2\rangle_{x(i\neq j)} = 1-\alpha$, and for small α: $\kappa_{i1(i>1)} \approx 1-\alpha/\gamma$. After replacement and keeping the terms to the first order of α, the components of $\boldsymbol{\mathcal{G}}$ and $\bar{\bar{\boldsymbol{\mathcal{G}}}}$ become:

$$\boldsymbol{\mathcal{G}}_{db}^{a,b} = \frac{\gamma}{8}\frac{x_a x_d}{s_d}(1 + \frac{s_d}{s_a}) \qquad\qquad \bar{\bar{\boldsymbol{\mathcal{G}}}}_{db,db}^{a,b} = \frac{\gamma^2}{64}\frac{\langle x_a^2 x_d^2\rangle_x}{s_d^2}(1 + \frac{s_d}{s_a})^2 = \frac{\gamma^2}{16}$$

$$\boldsymbol{\mathcal{G}}_{da}^{a,b} = \frac{-\gamma}{4}\frac{x_b x_d}{s_b}\frac{1+\kappa_{da}}{\sqrt{2(1+\kappa_{da}^2)}} \approx \frac{-\gamma}{4}\frac{x_b x_d}{s_d} \qquad \bar{\bar{\boldsymbol{\mathcal{G}}}}_{da,da}^{a,b} = \frac{\gamma^2}{16}\frac{\langle x_b^2 x_d^2\rangle_x}{s_d^2} = \frac{\gamma^2}{16}(1+\alpha)$$

$$\boldsymbol{\mathcal{G}}_{db}^{c,b} = \frac{\gamma}{4}\frac{x_c x_d}{s_d}, \qquad\qquad\qquad \bar{\bar{\boldsymbol{\mathcal{G}}}}_{db,db}^{c,b} = \frac{\gamma^2}{16}\frac{\langle x_c^2 x_d^2\rangle_x}{s_d^2} = \frac{\gamma^2}{16}(1+\alpha)$$

$$\boldsymbol{\mathcal{G}}_{dc}^{c,b} = \frac{-\gamma}{4}\frac{x_b x_d}{s_d}, \qquad\qquad\qquad \bar{\bar{\boldsymbol{\mathcal{G}}}}_{dc,dc}^{c,b} = \frac{\gamma^2}{16}\frac{\langle x_b^2 x_d^2\rangle_x}{s_d^2} = \frac{\gamma^2}{16}(1+\alpha)$$

Replacement of the above and the fluctuations from Eq.58 leads to Eq.27 of the main paper.

E.3. Calculation for large n and $\alpha \gg \gamma$

Here, we derive closed form solutions under the assumptions of $\gamma \ll 1$ and $s_a, s_b, s_a - s_b \gg \gamma$ (equivalently $\alpha \gg \gamma$). Using the second line of Eq.69, the pairwise diffusion coefficients become:

$$D_{ba} \approx \frac{n\eta^2}{2}(\langle\rho_{db}^2\rangle\bar{\bar{\boldsymbol{\mathcal{G}}}}_{db,db}^{a,b} + \langle\rho_{da}^2\rangle\bar{\bar{\boldsymbol{\mathcal{G}}}}_{da,da}^{a,b} + \langle\rho_{ad}^2\rangle\bar{\bar{\boldsymbol{\mathcal{G}}}}_{ad,ad}^{a,b} + 2\langle\rho_{ad}\rho_{da}\rangle\bar{\bar{\boldsymbol{\mathcal{G}}}}_{ad,da}^{a,b}) \qquad\qquad (n \gg 1)$$

$$\approx \frac{\eta^3\gamma^4 n}{2}\Big[\frac{\langle x_a^2 x_d^2\rangle\langle x_b^2 x_d^2\rangle(\frac{1}{s_a} + \frac{1}{s_b})^2}{64 s_b^3} + \frac{\langle x_a^2 x_d^2\rangle\langle x_b^2 x_d^2\rangle}{64 s_a^2 s_b^3} + \frac{\langle x_a^2 x_d^2\rangle\langle x_b^2 x_d^2\rangle}{64 s_a s_b^4} + \frac{\langle x_a^2 x_d^2\rangle\langle x_b^2 x_d^2\rangle}{16(s_a + s_b)s_a s_b^3}\Big] \qquad (\alpha \gg \gamma)$$

$$= \frac{\eta^3\gamma^4 n}{128}\frac{\langle x_a^2 x_d^2\rangle\langle x_b^2 x_d^2\rangle}{s_b^5}\Big[1 + \frac{3s_b}{s_a} + \frac{2s_b^2}{s_a^2} + \frac{4s_b^2}{s_a(s_a + s_b)}\Big]$$

$$= \frac{\eta^3\gamma^4 n}{16(1-\alpha)^3}\frac{1}{16}\Big[1 + 3(1-\alpha) + 2(1-\alpha)^2 + \frac{4(1-\alpha)^2}{2-\alpha}\Big] = \frac{\eta^3\gamma^4 n}{16(1-\alpha)^3}\Big[\frac{16 - 28\alpha + 15\alpha^2 - 2\alpha^3}{8(2-\alpha)}\Big]$$

$$D_{bc} \approx \frac{n\eta^2}{2}(\langle\rho_{db}^2\rangle\bar{\bar{\boldsymbol{\mathcal{G}}}}_{db,db}^{c,b} + \langle\rho_{dc}^2\rangle\bar{\bar{\boldsymbol{\mathcal{G}}}}_{dc,dc}^{c,b}) \qquad\qquad (n \gg 1)$$

$$= \frac{\eta^3\gamma^4 n}{16}\frac{\langle x_b^2 x_d^2\rangle\langle x_c^2 x_d^2\rangle}{s_b^5(1 - \frac{\gamma}{s_b})}$$

$$= \frac{\eta^3\gamma^4 n}{16(1-\alpha)^3(1 - \frac{\gamma}{s_b})} \overset{\gamma\ll 1}{\approx} \frac{\eta^3\gamma^4 n}{16(1-\alpha)^3} \qquad\qquad (70)$$

In the above, we replaced the appropriate quantities of $\langle\rho_{ij}\rho_{pq}\rangle$ and $\bar{\bar{\boldsymbol{\mathcal{G}}}}_{ij,pq}^{r,s}$, calculated from Eq.58 and Eq.15 respectively (see below). In deriving these terms, we used $\omega_i \approx 1$, $\langle x_i^2 x_j^2\rangle_{x(i\neq j)} = 1-\alpha$, $\lambda_{db} = 2(s_b - \gamma)$, $C_{db} = 1/(2\sqrt{1 - \gamma/s_b})$,

$\kappa_{ij(i>j>1)} = 1$ and $\kappa_{ad}, \kappa_{da} \approx 0$, all of which result from the mentioned assumptions.

$$\langle \rho_{db}^2 \rangle = \frac{\eta\gamma_r^2}{s_d^3}\langle x_b^2 x_d^2 \rangle = \langle \rho_{dc}^2 \rangle = \frac{\eta\gamma^2}{s_d^3}\langle x_c^2 x_d^2 \rangle = \frac{\eta\gamma^2}{(1-\alpha)^2}$$

$$\langle \rho_{da}^2 \rangle = \frac{\eta\gamma^2(\omega_a + \omega_d)}{2s_d[2 - \gamma(\frac{\kappa_{da}}{s_d} + \frac{1}{s_a})]}\frac{(\frac{\kappa_{da}}{s_d} + \frac{1}{s_a})^2}{1 + \kappa_{da}^2}\langle x_a^2 x_d^2 \rangle_x \overset{\alpha \gg \gamma}{\underset{\gamma \ll 1}{\approx}} \frac{\eta\gamma^2}{2s_a^2 s_d}\langle x_a^2 x_d^2 \rangle_x,$$

$$\langle \rho_{ad}^2 \rangle = \frac{\eta\gamma^2(\omega_a + \omega_d)}{2s_a[2 - \gamma(\frac{\kappa_{ad}}{s_a} + \frac{1}{s_d})]}\frac{(\frac{\kappa_{ad}}{s_a} + \frac{1}{s_d})^2}{1 + \kappa_{ad}^2}\langle x_a^2 x_d^2 \rangle_x \overset{\alpha \gg \gamma}{\underset{\gamma \ll 1}{\approx}} \frac{\eta\gamma^2 \langle x_a^2 x_d^2 \rangle_x}{2s_a s_d^2},$$

$$\langle \rho_{ad}\rho_{da} \rangle = \frac{\eta\gamma^2(\omega_a + \omega_d)}{[2s_a - \gamma(\kappa_{ad} + \frac{s_a}{s_d})] + [2s_d - \gamma(\kappa_{da} + \frac{s_d}{s_a})]}\frac{(\frac{\kappa_{ad}}{s_a} + \frac{1}{s_d})(\frac{\kappa_{da}}{s_d} + \frac{1}{s_a})}{\sqrt{1 + \kappa_{ad}^2}\sqrt{1 + \kappa_{da}^2}}\langle x_a^2 x_d^2 \rangle_x \overset{\alpha \gg \gamma}{\underset{\gamma \ll 1}{\approx}} \frac{\eta\gamma^2 \langle x_a^2 x_d^2 \rangle_x}{(s_a + s_d)s_a s_d}$$

$$\boldsymbol{G}_{db}^{a,b} = \frac{\gamma}{4}\frac{x_a x_d \sqrt{\omega_a}(\frac{1}{s_a} + \frac{1}{s_d})}{\omega_a + \omega_d} \qquad \bar{\bar{\boldsymbol{G}}}_{db,db}^{a,b} = \frac{\gamma^2}{16}\frac{\langle x_a^2 x_d^2 \rangle_x \omega_a(\frac{1}{s_a} + \frac{1}{s_d})^2}{(\omega_a + \omega_d)^2} \overset{\gamma \ll 1}{\approx} \frac{\gamma^2}{64}\frac{(\frac{s_d}{s_a} + 1)^2 \langle x_a^2 x_d^2 \rangle_x}{s_d^2}$$

$$\boldsymbol{G}_{da}^{a,b} = \frac{-\gamma}{2}\frac{x_b x_d \sqrt{\omega_a \omega_d}}{s_b(\omega_a + \omega_d)^{3/2}}\frac{1 + \kappa_{da}}{\sqrt{1 + \kappa_{da}^2}} \qquad \bar{\bar{\boldsymbol{G}}}_{da,da}^{a,b} = \frac{\gamma^2}{4}\frac{\langle x_b^2 x_d^2 \rangle_x \omega_a \omega_d}{s_b^2(\omega_a + \omega_d)^3}\frac{(1 + \kappa_{da})^2}{1 + \kappa_{da}^2} \overset{\alpha \gg \gamma}{\underset{\gamma \ll 1}{\approx}} \frac{\gamma^2}{32}\frac{\langle x_b^2 x_d^2 \rangle_x}{s_d^2}$$

$$\boldsymbol{G}_{ad}^{a,b} = \frac{-\gamma}{2}\frac{x_b x_d \sqrt{\omega_a \omega_d}}{s_b(\omega_a + \omega_d)^{3/2}}\frac{1 + \kappa_{ad}}{\sqrt{1 + \kappa_{ad}^2}} \qquad \bar{\bar{\boldsymbol{G}}}_{ad,ad}^{a,b} = \frac{\gamma^2}{4}\frac{\langle x_b^2 x_d^2 \rangle_x \omega_a \omega_d}{s_b^2(\omega_a + \omega_d)^3}\frac{(1 + \kappa_{ad})^2}{1 + \kappa_{ad}^2} \overset{\alpha \gg \gamma}{\underset{\gamma \ll 1}{\approx}} \frac{\gamma^2}{32}\frac{\langle x_b^2 x_d^2 \rangle_x}{s_d^2}$$

$$\bar{\bar{\boldsymbol{G}}}_{ad,da}^{a,b} = \frac{\gamma^2}{4}\frac{\langle x_b^2 x_d^2 \rangle_x \omega_a \omega_d}{s_b^2(\omega_a + \omega_d)^3}\frac{(1 + \kappa_{da})}{\sqrt{1 + \kappa_{da}^2}}\frac{(1 + \kappa_{ad})}{\sqrt{1 + \kappa_{ad}^2}} \overset{\alpha \gg \gamma}{\underset{\gamma \ll 1}{\approx}} \frac{\gamma^2}{32}\frac{\langle x_b^2 x_d^2 \rangle_x}{s_d^2}$$

$$\boldsymbol{G}_{db}^{c,b} = \frac{\gamma}{4}\frac{x_c x_d}{s_d \sqrt{\omega_d}}, \qquad \bar{\bar{\boldsymbol{G}}}_{db,db}^{c,b} = \frac{\gamma^2}{16}\frac{\langle x_c^2 x_d^2 \rangle_x}{s_d^2 \omega_d} = \frac{\gamma^2}{16}\frac{1}{(1-\alpha)\omega_d}$$

$$\boldsymbol{G}_{dc}^{c,b} = \frac{-\gamma}{4}\frac{x_b x_d}{s_d \sqrt{\omega_d}}, \qquad \bar{\bar{\boldsymbol{G}}}_{dc,dc}^{c,b} = \frac{\gamma^2}{16}\frac{\langle x_b^2 x_d^2 \rangle_x}{s_d^2 \omega_d} = \frac{\gamma^2}{16}\frac{1}{(1-\alpha)\omega_d}$$

Eq.30 in the main text can be derived from the above by using $D_a \approx nD_{ab}$ and $D_b \approx nD_{bc}$. These results reproduce the numerically measured curves of diffusion as a function of α very well.