A Appendix: Fukumizu Approach

For completeness, we reproduce the derivation from Fukumizu [1] of Equation 5. We consider the learning setting describe in section 2. Under the assumptions of equal input-output dimensions 2.1, whitened inputs 2.2 and zero-balanced weights 2.3, the weights dynamics yield

$$\tau \frac{d}{dt}\mathbf{W}_1 = \mathbf{W}_2^T(\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1\tilde{\mathbf{\Sigma}}^{xx}), \tag{16}$$

$$\tau \frac{d}{dt}\mathbf{W}_2 = (\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1\tilde{\mathbf{\Sigma}}^{xx})\mathbf{W}_1^T. \tag{17}$$

Under the assumption of whitened inputs 2.2, the dynamics simplify to

$$\tau \frac{d}{dt}\mathbf{W}_1 = \mathbf{W}_2^T(\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1), \tag{18}$$

$$\tau \frac{d}{dt}\mathbf{W}_2 = (\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1)\mathbf{W}_1^T. \tag{19}$$

We introduce the variables

$$\mathbf{Q} = \begin{bmatrix}\mathbf{W}_1^T \\ \mathbf{W}_2\end{bmatrix} \quad \text{and} \quad \mathbf{Q}\mathbf{Q}^T = \begin{bmatrix}\mathbf{W}_1^T\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{W}_2^T \\ \mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_2\mathbf{W}_2^T\end{bmatrix}. \tag{20}$$

We compute the time derivative

$$\tau \frac{d}{dt}(\mathbf{Q}\mathbf{Q}^T) = \tau \begin{bmatrix}\frac{d\mathbf{W}_1^T}{dt}\mathbf{W}_1 + \mathbf{W}_1^T\frac{d\mathbf{W}_1}{dt} & \frac{d\mathbf{W}_1^T}{dt}\mathbf{W}_2^T + \mathbf{W}_1^T\frac{d\mathbf{W}_2^T}{dt} \\ \frac{d\mathbf{W}_2}{dt}\mathbf{W}_1 + \mathbf{W}_2\frac{d\mathbf{W}_1}{dt} & \frac{d\mathbf{W}_2}{dt}\mathbf{W}_2^T + \mathbf{W}_2\frac{d\mathbf{W}_2^T}{dt}\end{bmatrix}. \tag{21}$$

Using equation 18 and 19 we compute the four quadrant separately giving

$$\tau(\frac{d\mathbf{W}_1^T}{dt}\mathbf{W}_1 + \mathbf{W}_1^T\frac{d\mathbf{W}_1}{dt}) = \tag{22}$$

$$= (\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1)^T\mathbf{W}_2\mathbf{W}_1 + \mathbf{W}_1^T\mathbf{W}_2^T(\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1) \tag{23}$$

$$= (\tilde{\mathbf{\Sigma}}^{yx})^T\mathbf{W}_2\mathbf{W}_1 + \mathbf{W}_1^T\mathbf{W}_2^T\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_1^T\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_1 - (\mathbf{W}_2\mathbf{W}_1)^T\mathbf{W}_2\mathbf{W}_1 \tag{24}$$

$$= (\tilde{\mathbf{\Sigma}}^{yx})^T\mathbf{W}_2\mathbf{W}_1 + \mathbf{W}_1^T\mathbf{W}_2^T\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_1^T\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_1 - \mathbf{W}_1^T\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_1, \tag{25}$$

$$\tau(\frac{d\mathbf{W}_1^T}{dt}\mathbf{W}_2^T + \mathbf{W}_1^T\frac{d\mathbf{W}_2^T}{dt}) = \tag{26}$$

$$= (\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1)^T\mathbf{W}_2\mathbf{W}_2^T + \mathbf{W}_1^T\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1)^T \tag{27}$$

$$= (\tilde{\mathbf{\Sigma}}^{yx})^T\mathbf{W}_2\mathbf{W}_2^T + \mathbf{W}_1^T\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx})^T - \mathbf{W}_1^T\mathbf{W}_1(\mathbf{W}_2\mathbf{W}_1)^T - (\mathbf{W}_2\mathbf{W}_1)^T\mathbf{W}_2\mathbf{W}_2^T, \tag{28}$$

$$= (\tilde{\mathbf{\Sigma}}^{yx})^T\mathbf{W}_2\mathbf{W}_2^T + \mathbf{W}_1^T\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx})^T - \mathbf{W}_1^T\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_2^T - \mathbf{W}_1^T\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_2^T, \tag{29}$$

$$\tag{30}$$

$$\tau(\frac{d\mathbf{W}_2}{dt}\mathbf{W}_1 + \mathbf{W}_2\frac{d\mathbf{W}_1}{dt}) \tag{31}$$

$$= (\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1)\mathbf{W}_1^T\mathbf{W}_1 + \mathbf{W}_2\mathbf{W}_2^T(\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1) \tag{32}$$

$$= \tilde{\mathbf{\Sigma}}^{yx}\mathbf{W}_1^T\mathbf{W}_1 + \mathbf{W}_2\mathbf{W}_2^T\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_1 - \mathbf{W}_2\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_1, \tag{33}$$

$$\tag{34}$$

$$\tau(\frac{d\mathbf{W}_2}{dt}\mathbf{W}_2^T + \mathbf{W}_2\frac{d\mathbf{W}_2^T}{dt}) = \tag{35}$$

$$= (\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1)\mathbf{W}_1^T\mathbf{W}_2^T + \mathbf{W}_2\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}_2\mathbf{W}_1)^T \tag{36}$$

$$= \tilde{\mathbf{\Sigma}}^{yx}\mathbf{W}_1^T\mathbf{W}_2^T + \mathbf{W}_2\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx})^T - \mathbf{W}_2\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_2^T - \mathbf{W}_2\mathbf{W}_1(\mathbf{W}_2\mathbf{W}_1)^T \tag{37}$$

$$= \tilde{\mathbf{\Sigma}}^{yx}\mathbf{W}_1^T\mathbf{W}_2^T + \mathbf{W}_2\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx})^T - \mathbf{W}_2\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_2^T - \mathbf{W}_2\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_2^T \tag{38}$$

$$= \tilde{\mathbf{\Sigma}}^{yx}\mathbf{W}_1^T\mathbf{W}_2^T + \mathbf{W}_2\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx})^T - \mathbf{W}_2\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_2^T - \mathbf{W}_2\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_2^T, \tag{39}$$

where we have used the assumption of zero-balanced weights 2.3 to simplify equation 25 and equation 39.

Defining

$$\mathbf{F} = \begin{bmatrix} 0 & (\tilde{\mathbf{\Sigma}}^{yx})^T \\ \tilde{\mathbf{\Sigma}}^{yx} & 0 \end{bmatrix}, \tag{40}$$

the gradient flow dynamics of $\mathbf{Q}\mathbf{Q}^T$(t) can be written as a differential matrix Riccati equation

$$\tau\frac{d}{dt}(\mathbf{Q}\mathbf{Q}^T) = \mathbf{F}\mathbf{Q}\mathbf{Q}^T + \mathbf{Q}\mathbf{Q}^T\mathbf{F} - (\mathbf{Q}\mathbf{Q}^T)^2. \tag{41}$$

We write $\tau\frac{d}{dt}(\mathbf{Q}\mathbf{Q}^T)$ for completeness

$$\tau\frac{d}{dt}(\mathbf{Q}\mathbf{Q}^T) =$$

$$\begin{bmatrix} 0 & (\tilde{\mathbf{\Sigma}}^{yx})^T \\ \tilde{\mathbf{\Sigma}}^{yx} & 0 \end{bmatrix}\begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{W}_2^T \\ \mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_2\mathbf{W}_2^T \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{W}_2^T \\ \mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_2\mathbf{W}_2^T \end{bmatrix}^T\begin{bmatrix} 0 & (\tilde{\mathbf{\Sigma}}^{yx})^T \\ \tilde{\mathbf{\Sigma}}^{yx} & 0 \end{bmatrix}$$
$$- \begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{W}_2^T \\ \mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_2\mathbf{W}_2^T \end{bmatrix}^2 \tag{42}$$

$$= \begin{bmatrix} 0 & (\tilde{\mathbf{\Sigma}}^{yx})^T \\ \tilde{\mathbf{\Sigma}}^{yx} & 0 \end{bmatrix}\begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{W}_2^T \\ \mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_2\mathbf{W}_2^T \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{W}_2^T \\ \mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_2\mathbf{W}_2^T \end{bmatrix}\begin{bmatrix} 0 & (\tilde{\mathbf{\Sigma}}^{yx})^T \\ \tilde{\mathbf{\Sigma}}^{yx} & 0 \end{bmatrix}$$
$$- \begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{W}_2^T \\ \mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_2\mathbf{W}_2^T \end{bmatrix}\begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{W}_2^T \\ \mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_2\mathbf{W}_2^T \end{bmatrix} \tag{43}$$

$$= \begin{bmatrix} (\tilde{\mathbf{\Sigma}}^{yx})^T\mathbf{W}_2\mathbf{W}_1 & (\tilde{\mathbf{\Sigma}}^{yx})^T\mathbf{W}_2\mathbf{W}_2^T \\ \tilde{\mathbf{\Sigma}}^{yx}\mathbf{W}_1^T\mathbf{W}_1 & \tilde{\mathbf{\Sigma}}^{yx}\mathbf{W}_1^T\mathbf{W}_2^T \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_2^T\tilde{\mathbf{\Sigma}}^{yx} & \mathbf{W}_1^T\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx})^T \\ \mathbf{W}_2\mathbf{W}_2^T\tilde{\mathbf{\Sigma}}^{yx} & \mathbf{W}_2\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx})^T \end{bmatrix}$$
$$- \begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{W}_2^T \\ \mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_2\mathbf{W}_2^T \end{bmatrix}\begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{W}_2^T \\ \mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_2\mathbf{W}_2^T \end{bmatrix} \tag{44}$$

$$= \begin{bmatrix} (\tilde{\mathbf{\Sigma}}^{yx})^T\mathbf{W}_2\mathbf{W}_1 & (\tilde{\mathbf{\Sigma}}^{yx})^T\mathbf{W}_2\mathbf{W}_2^T \\ \tilde{\mathbf{\Sigma}}^{yx}\mathbf{W}_1^T\mathbf{W}_1 & \tilde{\mathbf{\Sigma}}^{yx}\mathbf{W}_1^T\mathbf{W}_2^T \end{bmatrix} + \begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_2^T\tilde{\mathbf{\Sigma}}^{yx} & \mathbf{W}_1^T\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx})^T \\ \mathbf{W}_2\mathbf{W}_2^T\tilde{\mathbf{\Sigma}}^{yx} & \mathbf{W}_2\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx})^T \end{bmatrix}$$
$$- \begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_1 + \mathbf{W}_1^T\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_1^T\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_2^T + \mathbf{W}_1^T\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_2^T \\ \mathbf{W}_2\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_1 + \mathbf{W}_2\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_1 & \mathbf{W}_2\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_2^T + \mathbf{W}_2\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_2^T \end{bmatrix} \tag{45}$$

$$= \begin{bmatrix} (\tilde{\mathbf{\Sigma}}^{yx})^T\mathbf{W}_2\mathbf{W}_1 + \mathbf{W}_1^T\mathbf{W}_2^T\tilde{\mathbf{\Sigma}}^{yx} & \\ -\mathbf{W}_1^T\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_1 - \mathbf{W}_1^T\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_1 & (\tilde{\mathbf{\Sigma}}^{yx})^T\mathbf{W}_2\mathbf{W}_2^T + \mathbf{W}_1^T\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx})^T \\ & -\mathbf{W}_1^T\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_2^T - \mathbf{W}_1^T\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_2^T \\ \tilde{\mathbf{\Sigma}}^{yx}\mathbf{W}_1^T\mathbf{W}_1 + \mathbf{W}_2\mathbf{W}_2^T\tilde{\mathbf{\Sigma}}^{yx} & \\ -\mathbf{W}_2\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_1 - \mathbf{W}_2\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_1 & \tilde{\mathbf{\Sigma}}^{yx}\mathbf{W}_1^T\mathbf{W}_2^T + \mathbf{W}_2\mathbf{W}_1(\tilde{\mathbf{\Sigma}}^{yx})^T \\ & -\mathbf{W}_2\mathbf{W}_1\mathbf{W}_1^T\mathbf{W}_2^T - \mathbf{W}_2\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_2^T \end{bmatrix} \tag{46}$$

□

The four quadrant of 46 are equivalent to equations 25,29,33 and 39 respectively.

Assuming that $\mathbf{Q}(0)$ is full rank, the continuous differential equation 41 has a unique solution for all $t \geq 0$

$$\mathbf{Q}\mathbf{Q}^T(t) = e^{\mathbf{F}\frac{t}{\tau}}\mathbf{Q}(0)\left[\mathbf{I} + \frac{1}{2}\mathbf{Q}(0)^T\left(e^{\mathbf{F}\frac{t}{\tau}}\mathbf{F}^{-1}e^{\mathbf{F}\frac{t}{\tau}} - \mathbf{F}^{-1}\right)\mathbf{Q}(0)\right]^{-1}\mathbf{Q}(0)^T e^{\mathbf{F}\frac{t}{\tau}}. \tag{47}$$

B Appendix: Network's internal representations

B.1 Representational similarity analysis

The task-relevant representational similarity matrix [50] of the hidden layer, calculated from the inputs $\mathbf{H} = \mathbf{W}_1\mathbf{X}$ is

$$\begin{aligned}\mathrm{RSM}_I(t) &= \mathbf{H}^T(t)\mathbf{H}(t) && (48)\\ &= (\mathbf{W}_1(t)\mathbf{X})^T\mathbf{W}_1(t)\mathbf{X} && (49)\\ &= \mathbf{X}^T(\mathbf{W}_1^T\mathbf{W}_1)(t)\mathbf{X}. && (50)\end{aligned}$$

Similarly, the representational similarity matrix of the hidden layer, calculated from the outputs $\tilde{\mathbf{H}} = \mathbf{W}_2^+ Y$, where $+$ denotes the pseudoinverse, is

$$\begin{aligned}\mathrm{RSM}_O(t) &= \tilde{\mathbf{H}}^T(t)\tilde{\mathbf{H}}(t) && (51)\\ &= (\mathbf{W}_2^+(t)Y)^T\mathbf{W}_2^+(t)Y && (52)\\ &= Y^T(\mathbf{W}_2\mathbf{W}_2^T(t))^+Y. && (53)\end{aligned}$$

B.2 Finite-width neural tangent kernel

In the following, we derive the finite-width neural tangent kernel [39] for a two-layer linear network. Starting with the network function at time t

$$F_t(\mathbf{X}) = \mathbf{W}_2\mathbf{W}_1\mathbf{X}, \tag{54}$$

the discrete time gradient descent dynamics of the next time step yields

$$\begin{aligned}F_{t+1}(\mathbf{X}) &= \left(\mathbf{W}_2 - \eta\frac{\partial\mathcal{L}}{\partial\mathbf{W}_2}\right)\left(\mathbf{W}_1 - \eta\frac{\partial\mathcal{L}}{\partial\mathbf{W}_1}\right)\mathbf{X} && (55)\\ &= \mathbf{W}_2\mathbf{W}_1\mathbf{X} - \eta\left(\mathbf{W}_2\frac{\partial\mathcal{L}}{\partial\mathbf{W}_1} + \frac{\partial\mathcal{L}}{\partial\mathbf{W}_2}\mathbf{W}_1 - \eta\frac{\partial\mathcal{L}}{\partial\mathbf{W}_2}\frac{\partial\mathcal{L}}{\partial\mathbf{W}_1}\right)\mathbf{X}. && (56)\end{aligned}$$

The network function's gradient flow can then be derived as

$$\begin{aligned}\frac{F_{t+1}(\mathbf{X}) - F_t(\mathbf{X})}{\eta} &= -\left(\mathbf{W}_2\frac{\partial\mathcal{L}}{\partial\mathbf{W}_1} + \frac{\partial\mathcal{L}}{\partial\mathbf{W}_2}\mathbf{W}_1 - \eta\frac{\partial\mathcal{L}}{\partial\mathbf{W}_2}\frac{\partial\mathcal{L}}{\partial\mathbf{W}_1}\right)\mathbf{X} && (57)\\ \xrightarrow[\eta\to 0]{} \frac{d}{dt}F(\mathbf{X}) &= -\left(\mathbf{W}_2\frac{\partial\mathcal{L}}{\partial\mathbf{W}_1} + \frac{\partial\mathcal{L}}{\partial\mathbf{W}_2}\mathbf{W}_1\right)\mathbf{X}. && (58)\end{aligned}$$

Substituting the partial derivatives

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{W}_1} &= \frac{1}{2}\frac{\partial}{\partial\mathbf{W}_1}||\mathbf{W}_2\mathbf{W}_1\mathbf{X} - \mathbf{Y}||_F^2 && (59)\\ &= \mathbf{W}_2^T(\mathbf{W}_2\mathbf{W}_1\mathbf{X} - \mathbf{Y})\mathbf{X}^T && (60)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{W}_2} &= \frac{1}{2}\frac{\partial}{\partial\mathbf{W}_2}||\mathbf{W}_2\mathbf{W}_1\mathbf{X} - \mathbf{Y}||_F^2 && (61)\\ &= (\mathbf{W}_2\mathbf{W}_1\mathbf{X} - \mathbf{Y})\mathbf{X}^T\mathbf{W}_1^T && (62)\end{aligned}$$

then yields

$$\frac{d}{dt}F(X) = -\mathbf{W}_2\mathbf{W}_2^T(\mathbf{W}_2\mathbf{W}_1\mathbf{X} - \mathbf{Y})\mathbf{X}^T\mathbf{X} - (\mathbf{W}_2\mathbf{W}_1\mathbf{X} - \mathbf{Y})\mathbf{X}^T\mathbf{W}_1^T\mathbf{W}_1\mathbf{X}. \tag{63}$$

Finally, we introduce the identity matrix $\mathbf{I}_{N_o}$ of size N_o and apply row-wise vectoriasation $\mathrm{vec_r}(F(\mathbf{X})) := f(\mathbf{X})$ and the identity $\mathrm{vec_r}(ABC) = (A \otimes C^T)\,\mathrm{vec_r}(B)$ to derive the neural

tangent kernel

$$\frac{d}{dt}F(X) = -\mathbf{W}_2\mathbf{W}_2^T(\mathbf{W}_2\mathbf{W}_1\mathbf{X}-\mathbf{Y})\mathbf{X}^T\mathbf{X} - \mathbf{I}_{N_o}(\mathbf{W}_2\mathbf{W}_1\mathbf{X}-\mathbf{Y})\mathbf{X}^T\mathbf{W}_1^T\mathbf{W}_1\mathbf{X} \tag{64}$$

$$\Leftrightarrow \frac{d}{dt}f(\mathbf{X}) = -\left(\underbrace{\mathbf{W}_2\mathbf{W}_2^T\otimes\mathbf{X}^T\mathbf{X} + \mathbf{I}\otimes\mathbf{X}^T\mathbf{W}_1^T\mathbf{W}_1\mathbf{X}}_{\text{NTK}}\right)\text{vec}_\text{r}(\mathbf{W}_2\mathbf{W}_1\mathbf{X}-\mathbf{Y}) \tag{65}$$

$$= -\left(\left[\mathbf{W}_2\otimes\mathbf{X}^T, \mathbf{I}\otimes\mathbf{X}^T\mathbf{W}_1^T\right]\left[\mathbf{W}_2\otimes\mathbf{X}^T, \mathbf{I}\otimes\mathbf{X}^T\mathbf{W}_1^T\right]^T\right)\text{vec}_\text{r}\left(\frac{\partial\mathcal{L}}{\partial F}\right) \tag{66}$$

$$= -\left(\left[\nabla_{\mathbf{W}_1}f, \nabla_{\mathbf{W}_2}f\right]\left[\nabla_{\mathbf{W}_1}f, \nabla_{\mathbf{W}_2}f\right]^T\right)\frac{\partial\mathcal{L}}{\partial f} \tag{67}$$

$$= -\left(\nabla_\theta f\nabla_\theta f^T\right)\frac{\partial\mathcal{L}}{\partial f}, \tag{68}$$

where $[A, B]$ denotes concatenation.

C Appendix: Exact learning dynamics with prior knowledge

C.1 Proof of Theorem 3.1

In the following, we prove that Equation 11 is in fact a solution to the matrix Riccati equation arising from gradient flow (Equation 41). We prove the theorem by directly substituting our solution for $\mathbf{Q}\mathbf{Q}^T(t)$ into the matrix Riccati equation.

C.1.1 Unequal input-output dimension

We start with the following equation

$$\begin{aligned}\mathbf{Q}\mathbf{Q}^T(t) = &\underbrace{\left[\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T + 2\mathbf{M}\mathbf{M}^T\right]\mathbf{Q}(0)}_{\mathbf{L}} \\ &\underbrace{\left[\mathbf{I} + \frac{1}{2}\mathbf{Q}(0)^T\left(\mathbf{O}\left(e^{2\mathbf{\Lambda}\frac{t}{\tau}}-\mathbf{I}\right)\mathbf{\Lambda}^{-1}\mathbf{O}^T + 4\frac{t}{\tau}\mathbf{M}\mathbf{M}^T\right)\mathbf{Q}(0)\right]^{-1}}_{\mathbf{C}^{-1}} \\ &\underbrace{\mathbf{Q}(0)^T\left[\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T + 2\mathbf{M}\mathbf{M}^T\right]}_{\mathbf{R}} \\ =&\mathbf{L}\mathbf{C}^{-1}\mathbf{R},\end{aligned} \tag{69, 70}$$

which is identical to Equation 11 in the main text, as we verify in Section C.2 (by reversing the derivation from Equation 154 to Equation 130). Substituting our solution into the matrix Riccati equation then yields

$$\tau\frac{d}{dt}\mathbf{Q}\mathbf{Q}^T = \mathbf{F}\mathbf{Q}\mathbf{Q}^T + \mathbf{Q}\mathbf{Q}^T\mathbf{F} - (\mathbf{Q}\mathbf{Q}^T)^2 \tag{71}$$

$$\Rightarrow \tau\frac{d}{dt}\mathbf{L}\mathbf{C}^{-1}\mathbf{R} \stackrel{?}{=} \mathbf{F}\mathbf{L}\mathbf{C}^{-1}\mathbf{R} + \mathbf{L}\mathbf{C}^{-1}\mathbf{R}\mathbf{F} - \mathbf{L}\mathbf{C}^{-1}\mathbf{R}\mathbf{L}\mathbf{C}^{-1}\mathbf{R}. \tag{72}$$

Next, we note that

$$\mathbf{O}^T\mathbf{O} = \frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}} & \tilde{\mathbf{V}}\\ \tilde{\mathbf{U}} & -\tilde{\mathbf{U}}\end{bmatrix}^T\frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}} & \tilde{\mathbf{V}}\\ \tilde{\mathbf{U}} & -\tilde{\mathbf{U}}\end{bmatrix} = \mathbf{I}, \tag{73}$$

$$\mathbf{O}^T\mathbf{M} = \frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}}^T & \tilde{\mathbf{U}}^T\\ \tilde{\mathbf{V}}^T & -\tilde{\mathbf{U}}^T\end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}}_\perp\\ \tilde{\mathbf{U}}_\perp\end{bmatrix} \tag{74}$$

$$= \frac{1}{2}\begin{bmatrix}\tilde{\mathbf{V}}^T\tilde{\mathbf{V}}_\perp + \tilde{\mathbf{U}}^T\tilde{\mathbf{U}}_\perp\\ \tilde{\mathbf{V}}^T\tilde{\mathbf{V}}_\perp - \tilde{\mathbf{U}}^T\tilde{\mathbf{U}}_\perp\end{bmatrix} \tag{75}$$

$$= \mathbf{0} \tag{76}$$

and

$$\mathbf{M}^T\mathbf{O} = \frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}}_\perp^T & \tilde{\mathbf{U}}_\perp^T\end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}} & \tilde{\mathbf{V}}\\ \tilde{\mathbf{U}} & -\tilde{\mathbf{U}}\end{bmatrix} \tag{77}$$

$$= \frac{1}{2}\begin{bmatrix}\tilde{\mathbf{V}}_\perp^T\tilde{\mathbf{V}} + \tilde{\mathbf{U}}_\perp^T\tilde{\mathbf{U}}\\ \tilde{\mathbf{V}}_\perp^T\tilde{\mathbf{V}} - \tilde{\mathbf{U}}_\perp^T\tilde{\mathbf{U}}\end{bmatrix} \tag{78}$$

$$= \mathbf{0}. \tag{79}$$

Then, using the chain rule $\partial(\mathbf{AB}) = (\partial\mathbf{A})\mathbf{B} + \mathbf{A}(\partial\mathbf{B})$ and the identities

$$\frac{d}{dt}(\mathbf{A}^{-1}) = \mathbf{A}^{-1}(\frac{d}{dt}\mathbf{A})\mathbf{A}^{-1} \qquad \text{and} \qquad \frac{d}{dt}(e^{t\mathbf{A}}) = \mathbf{A}e^{t\mathbf{A}} = e^{t\mathbf{A}}\mathbf{A} \tag{80}$$

we get

$$\tau\frac{d}{dt}\mathbf{Q}\mathbf{Q}^T = \tau\frac{d}{dt}\left(\mathbf{L}\mathbf{C}^{-1}\mathbf{R}\right) \tag{81}$$

$$= \tau\left(\frac{d}{dt}\mathbf{L}\right)\mathbf{C}^{-1}\mathbf{R} + \tau\mathbf{L}\left(\frac{d}{dt}C^{-1}\mathbf{R}\right) \tag{82}$$

$$= \tau\left(\frac{d}{dt}\mathbf{L}\right)\mathbf{C}^{-1}\mathbf{R} + \tau\mathbf{L}\mathbf{C}^{-1}\left(\frac{d}{dt}\mathbf{R}\right) + \tau\mathbf{L}\left(\frac{d}{dt}\mathbf{C}^{-1}\right)\mathbf{R}, \tag{83}$$

with

$$\tau\left(\frac{d}{dt}\mathbf{L}\right)\mathbf{C}^{-1}\mathbf{R} = \tau\mathbf{O}\frac{1}{\tau}\mathbf{\Lambda}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{Q}(0)\mathbf{C}^{-1}\mathbf{R} \tag{84}$$

$$= \mathbf{O}\mathbf{\Lambda}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{Q}(0)\mathbf{C}^{-1}\mathbf{R} \tag{85}$$

$$= \left[\mathbf{O}\mathbf{\Lambda}\mathbf{O}^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{Q}(0) + 2\mathbf{O}\mathbf{\Lambda}\underbrace{\mathbf{O}^T\mathbf{M}}_{\mathbf{0}}\mathbf{M}^T\mathbf{Q}(0)\right]\mathbf{C}^{-1}\mathbf{R} \tag{86}$$

$$= \mathbf{F}\mathbf{L}\mathbf{C}^{-1}\mathbf{R}, \tag{87}$$

$$\tau\mathbf{L}\mathbf{C}^{-1}\left(\frac{d}{dt}\mathbf{R}\right) = \tau\mathbf{L}\mathbf{C}^{-1}\mathbf{Q}(0)^T\mathbf{O}\frac{1}{\tau}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{\Lambda}\mathbf{O}^T \tag{88}$$

$$= \mathbf{L}\mathbf{C}^{-1}\mathbf{Q}(0)^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{\Lambda}\mathbf{O}^T \tag{89}$$

$$= \mathbf{L}\mathbf{C}^{-1}\left[\mathbf{Q}(0)^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{O}\mathbf{\Lambda}\mathbf{O}^T + 2\mathbf{Q}(0)^T\mathbf{M}\underbrace{\mathbf{M}^T\mathbf{O}}_{\mathbf{0}}\mathbf{\Lambda}\mathbf{O}^T\right] \tag{90}$$

$$= \mathbf{L}\mathbf{C}^{-1}\mathbf{R}\mathbf{F} \tag{91}$$

and

$$\tau\mathbf{L}\left(\frac{d}{dt}\mathbf{C}^{-1}\right)\mathbf{R} = -\tau\mathbf{L}\mathbf{C}^{-1}\left(\frac{d}{dt}\mathbf{C}\right)\mathbf{C}^{-1}\mathbf{R} \tag{92}$$

$$= -\mathbf{L}\mathbf{C}^{-1}\Bigg[\tau\frac{1}{2}\mathbf{Q}(0)^T\mathbf{O}2\frac{1}{\tau}e^{2\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{\Lambda}\mathbf{\Lambda}^{-1}\mathbf{O}^T\mathbf{Q}(0) \tag{93}$$

$$+ \tau\frac{1}{2}\mathbf{Q}(0)^T4\frac{1}{\tau}\mathbf{M}\mathbf{M}^T\mathbf{Q}(0)\Bigg]\mathbf{C}^{-1}\mathbf{R}$$

$$= -\mathbf{L}\mathbf{C}^{-1}\Big[\mathbf{Q}(0)^T\mathbf{O}e^{2\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{Q}(0) + 2\mathbf{Q}(0)^T\mathbf{M}\mathbf{M}^T\mathbf{Q}(0)\Big]\mathbf{C}^{-1}\mathbf{R} \tag{94}$$

$$= -\mathbf{L}\mathbf{C}^{-1}\Bigg[\mathbf{Q}(0)^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{Q}(0)$$

$$+ 2\mathbf{Q}(0)^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\underbrace{\mathbf{O}^T\mathbf{M}}_{\mathbf{0}}\mathbf{M}^T\mathbf{Q}(0) \tag{95}$$

$$+ 2\mathbf{Q}(0)^T\mathbf{M}\underbrace{\mathbf{M}^T\mathbf{O}}_{\mathbf{0}}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{Q}(0)$$

$$+ 4\mathbf{Q}(0)^T\mathbf{M}\mathbf{M}^T\mathbf{M}\mathbf{M}^T\mathbf{Q}(0)\Bigg]\mathbf{C}^{-1}\mathbf{R}$$

$$= -\mathbf{L}\mathbf{C}^{-1}\mathbf{R}\mathbf{L}\mathbf{C}^{-1}\mathbf{R}. \tag{96}$$

Finally, substituting Equations 84, 88 and 92 into the left hand side of Equation 72 proves equality.

□

C.1.2 Equal input-output dimension

In the case of equal input-output dimensions $\tilde{\mathbf{U}}_\perp = \tilde{\mathbf{V}}_\perp = 0$ Equation 69 reduces to

$$\begin{aligned}\mathbf{Q}\mathbf{Q}^T(t) = &\underbrace{\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{Q}(0)}_{\mathbf{L}} \\ &\underbrace{\left[\mathbf{I} + \frac{1}{2}\mathbf{Q}(0)^T\mathbf{O}e^{2\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{\Lambda}^{-1}\mathbf{O}^T\mathbf{Q}(0) - \frac{1}{2}\mathbf{Q}(0)^T\mathbf{O}\mathbf{\Lambda}^{-1}\mathbf{O}^T\mathbf{Q}(0)\right]^{-1}}_{\mathbf{C}^{-1}} \qquad (97) \\ &\underbrace{\mathbf{Q}(0)^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T}_{\mathbf{R}} \\ = &\mathbf{L}\mathbf{C}^{-1}\mathbf{R}. \qquad (98)\end{aligned}$$

Therefore, analogously to the proof for unequal input-output dimensions, it follows that

$$\begin{aligned}\tau\frac{d}{dt}\mathbf{Q}\mathbf{Q}^T &= \tau\frac{d}{dt}\mathbf{L}\mathbf{C}^{-1}\mathbf{R} \qquad (99)\\ &= \tau\left(\frac{d}{dt}\mathbf{L}\right)\mathbf{C}^{-1}\mathbf{R} + \tau\mathbf{L}\left(\frac{d}{dt}C^{-1}\mathbf{R}\right) \qquad (100)\\ &= \tau\left(\frac{d}{dt}\mathbf{L}\right)\mathbf{C}^{-1}\mathbf{R} + \tau\mathbf{L}\mathbf{C}^{-1}\left(\frac{d}{dt}\mathbf{R}\right) + \tau\mathbf{L}\left(\frac{d}{dt}\mathbf{C}^{-1}\right)\mathbf{R}, \qquad (101)\end{aligned}$$

with

$$\begin{aligned}\tau\left(\frac{d}{dt}\mathbf{L}\right)\mathbf{C}^{-1}\mathbf{R} &= \tau\mathbf{O}\mathbf{\Lambda}\frac{1}{\tau}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{Q}(0)\mathbf{C}^{-1}\mathbf{R} \qquad (102)\\ &= \mathbf{O}\mathbf{\Lambda}\mathbf{O}^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{Q}(0)\mathbf{C}^{-1}\mathbf{R} \qquad (103)\\ &= \mathbf{F}\mathbf{L}\mathbf{C}^{-1}\mathbf{R}, \qquad (104)\end{aligned}$$

$$\begin{aligned}\tau\mathbf{L}\mathbf{C}^{-1}\left(\frac{d}{dt}\mathbf{R}\right) &= \tau\mathbf{L}\mathbf{C}^{-1}\mathbf{Q}(0)^T\mathbf{O}\frac{1}{\tau}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{\Lambda}\mathbf{O}^T \qquad (105)\\ &= \mathbf{L}\mathbf{C}^{-1}\mathbf{Q}(0)^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{O}\mathbf{\Lambda}\mathbf{O}^T \qquad (106)\\ &= \mathbf{L}\mathbf{C}^{-1}\mathbf{R}\mathbf{F}, \qquad (107)\end{aligned}$$

and

$$\begin{aligned}\tau\mathbf{L}\left(\frac{d}{dt}\mathbf{C}^{-1}\mathbf{R}\right) &= -\tau\mathbf{L}\mathbf{C}^{-1}\left(\frac{d}{dt}\mathbf{C}\right)\mathbf{C}^{-1}\mathbf{R} \qquad (108)\\ &= -\tau\mathbf{L}\mathbf{C}^{-1}\left(\frac{1}{2}\mathbf{Q}(0)^T\mathbf{O}e^{2\mathbf{\Lambda}\frac{t}{\tau}}\frac{2}{\tau}\mathbf{\Lambda}\mathbf{\Lambda}^{-1}\mathbf{O}^T\mathbf{Q}(0)\right)\mathbf{C}^{-1}\mathbf{R} \qquad (109)\\ &= -\tau\mathbf{L}\mathbf{C}^{-1}\mathbf{Q}(0)^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{Q}(0)\mathbf{C}^{-1}\mathbf{R} \qquad (110)\\ &= -\mathbf{L}\mathbf{C}^{-1}\mathbf{R}\mathbf{L}\mathbf{C}^{-1}\mathbf{R}. \qquad (111)\end{aligned}$$

Finally, substituting Equations 102, 105 and 108 into the left hand side of Equation 72 proves equality.

□

C.2 Derivation of the exact learning dynamics

In the following, we outline how the solution to the matrix Ricatti equation can be acquired. Let the input and output dimension of a two-layer linear network (equation 1) be denoted by N_i and N_o respectively. Further, let $N_m = \min(N_i, N_o)$ denote the smaller one of the two. The compact

singular value decomposition of the initial network function and the input-output correlation of the task is then

$$\text{SVD}(\mathbf{W}_2(0)\mathbf{W}_1(0)) = \mathbf{U}\mathbf{S}\mathbf{V}^T \quad \text{and} \quad \text{SVD}(\tilde{\mathbf{\Sigma}}^{yx}) = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T. \tag{112}$$

Here, $\mathbf{U}$ and $\tilde{\mathbf{U}} \in \mathbb{R}^{N_o \times N_m}$ denote the left singular vectors, $\mathbf{S}$ and $\tilde{\mathbf{S}} \in \mathbb{R}^{N_m \times N_m}$ the square matrix with ordered, non-zero eigenvalues on its diagonal and $\mathbf{V}$ and $\tilde{\mathbf{V}} \in \mathbb{R}^{N_i \times N_m}$ the corresponding right singular vectors. Please note that when using compact singular value decomposition, in the case of unequal input-output dimensions ($N_i \neq N_o$) the right and left singular vectors are not generally square and orthonormal.

More specifically, in the case of $N_i < N_o$, $\tilde{\mathbf{U}}^T\tilde{\mathbf{U}} = \tilde{\mathbf{V}}^T\tilde{\mathbf{V}} = \tilde{\mathbf{V}}\tilde{\mathbf{V}}^T = \mathbf{I} \in \mathbb{R}^{N_i \times N_i}$ but $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T \neq \mathbf{I} \in \mathbb{R}^{N_o \times N_o}$. In this case, we use $\tilde{\mathbf{U}}_\perp \in \mathbb{R}^{N_o \times (N_o - N_i)}$ to denote the matrix that contains orthogonal column vectors such that the concatenation $\begin{bmatrix}\tilde{\mathbf{U}} & \tilde{\mathbf{U}}_\perp\end{bmatrix}$ is orthonormal and $\tilde{\mathbf{V}}_\perp \in \mathbb{R}^{N_i \times (N_o - N_i)}$ to denote a matrix of zeros.

Conversely, in the case of $N_i > N_o$, $\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T = \tilde{\mathbf{U}}^T\tilde{\mathbf{U}} = \tilde{\mathbf{V}}^T\tilde{\mathbf{V}} = \mathbf{I} \in \mathbb{R}^{N_o \times N_o}$ but $\tilde{\mathbf{V}}^T\tilde{\mathbf{V}} \neq \mathbf{I} \in \mathbb{R}^{N_i \times N_i}$ and we define $\tilde{\mathbf{V}}_\perp \in \mathbb{R}^{N_i \times (N_i - N_o)}$ such that $\begin{bmatrix}\tilde{\mathbf{V}} & \tilde{\mathbf{V}}_\perp\end{bmatrix}$ is orthonormal and $\tilde{\mathbf{U}}_\perp \in \mathbb{R}^{N_o \times (N_o - N_i)}$ to denote a matrix of zeros.

C.2.1 Inverse and matrix exponential of F

The solution to the matrix Riccati equation as provided by Fukumizu [1] requires calculation of the inverse $\mathbf{F}^{-1}$ and the matrix exponential $e^{\mathbf{F}\frac{t}{\tau}}$. To this end, we diagonalise $\mathbf{F}$ by completing its basis by incorporating zero eigenvalues as illustrated below

$$\begin{aligned}
\mathbf{F} &= \begin{bmatrix} 0 & \tilde{\mathbf{V}}\tilde{\mathbf{S}}\tilde{\mathbf{U}}^T \\ \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T & 0 \end{bmatrix} && (113)\\
&= \frac{1}{\sqrt{2}}\begin{bmatrix} \tilde{\mathbf{V}} & \tilde{\mathbf{V}} & \sqrt{2}\tilde{\mathbf{V}}_\perp \\ \tilde{\mathbf{U}} & -\tilde{\mathbf{U}} & \sqrt{2}\tilde{\mathbf{U}}_\perp \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{S}} & 0 & 0 \\ 0 & -\tilde{\mathbf{S}} & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{1}{\sqrt{2}}\begin{bmatrix} \tilde{\mathbf{V}} & \tilde{\mathbf{V}} & \sqrt{2}\tilde{\mathbf{V}}_\perp \\ \tilde{\mathbf{U}} & -\tilde{\mathbf{U}} & \sqrt{2}\tilde{\mathbf{U}}_\perp \end{bmatrix}^T && (114)\\
&= \mathbf{P}\mathbf{\Gamma}\mathbf{P}^T. && (115)
\end{aligned}$$

Note that $\mathbf{P}^T\mathbf{P} = \mathbf{P}\mathbf{P}^T = \mathbf{I}$ and therefore $\mathbf{P}^T = \mathbf{P}^{-1}$. We then use the diagonalisation of $\mathbf{F}$ to rewrite the matrix exponential

$$\begin{aligned}
e^{\mathbf{F}\frac{t}{\tau}} &= \mathbf{P}e^{\mathbf{\Gamma}}\mathbf{P}^T && (116)\\
&= \frac{1}{\sqrt{2}}\begin{bmatrix} \tilde{\mathbf{V}} & \tilde{\mathbf{V}} & \sqrt{2}\mathbf{V}_\perp \\ \tilde{\mathbf{U}} & -\tilde{\mathbf{U}} & \sqrt{2}\mathbf{U}_\perp \end{bmatrix} \begin{bmatrix} e^{\tilde{\mathbf{S}}\frac{t}{\tau}} & 0 & 0 \\ 0 & e^{-\tilde{\mathbf{S}}\frac{t}{\tau}} & 0 \\ 0 & 0 & e^{0} \end{bmatrix} \frac{1}{\sqrt{2}}\begin{bmatrix} \tilde{\mathbf{V}} & \tilde{\mathbf{V}} & \sqrt{2}\mathbf{V}_\perp \\ \tilde{\mathbf{U}} & -\tilde{\mathbf{U}} & \sqrt{2}\mathbf{U}_\perp \end{bmatrix}^T && (117)\\
&= \frac{1}{2}\begin{bmatrix} \tilde{\mathbf{V}}e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{V}}^T + \tilde{\mathbf{V}}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{V}}^T + 2\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T & \tilde{\mathbf{V}}e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{U}}^T - \tilde{\mathbf{V}}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{U}}^T + 2\tilde{\mathbf{V}}_\perp\tilde{\mathbf{U}}_\perp^T \\ \tilde{\mathbf{U}}e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{V}}^T - \tilde{\mathbf{U}}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{V}}^T + 2\tilde{\mathbf{U}}_\perp\tilde{\mathbf{V}}_\perp^T & \tilde{\mathbf{U}}e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{U}}^T - \tilde{\mathbf{U}}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{U}}^T + 2\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T \end{bmatrix} && (118)\\
&= \frac{1}{\sqrt{2}}\begin{bmatrix} \tilde{\mathbf{V}} & \tilde{\mathbf{V}} \\ \tilde{\mathbf{U}} & -\tilde{\mathbf{U}} \end{bmatrix} \begin{bmatrix} e^{\tilde{\mathbf{S}}\frac{t}{\tau}} & 0 \\ 0 & e^{-\tilde{\mathbf{S}}\frac{t}{\tau}} \end{bmatrix} \frac{1}{\sqrt{2}}\begin{bmatrix} \tilde{\mathbf{V}} & \tilde{\mathbf{V}} \\ \tilde{\mathbf{U}} & -\tilde{\mathbf{U}} \end{bmatrix}^T + 2\frac{1}{\sqrt{2}}\begin{bmatrix} \tilde{\mathbf{V}}_\perp \\ \tilde{\mathbf{U}}_\perp \end{bmatrix} \frac{1}{\sqrt{2}}\begin{bmatrix} \tilde{\mathbf{V}}_\perp \\ \tilde{\mathbf{U}}_\perp \end{bmatrix}^T && (119)\\
&= \mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O} + 2\mathbf{M}\mathbf{M}^T. && (120)
\end{aligned}$$

As the inverse $\mathbf{F}^{-1} = \mathbf{P}\mathbf{\Gamma}^{-1}\mathbf{P}^T$ is not well defined for a $\mathbf{\Gamma}$ with zero eigenvalues. We study eigenvalues of value zero by analysing the limiting behaviour of

$$e^{\mathbf{F}\frac{t}{\tau}}\mathbf{F}^{-1}e^{\mathbf{F}\frac{t}{\tau}} - \mathbf{F}^{-1} \tag{121}$$

for a single mode

$$\lim_{\epsilon\to 0}\left[e^{\frac{\epsilon t}{\tau}}\frac{1}{\epsilon}e^{\frac{\epsilon t}{\tau}}-\frac{1}{\epsilon}\right]=\lim_{\epsilon\to 0}\left[\frac{e^{\frac{2\epsilon t}{\tau}}-1}{\epsilon}\right] \tag{122}$$

$$\xrightarrow{\text{L'Hospital}}\lim_{\epsilon\to 0}\left[\frac{\frac{\partial}{\partial\epsilon}\left(e^{\frac{2\epsilon t}{\tau}}-1\right)}{\frac{\partial}{\partial\epsilon}\epsilon}\right] \tag{123}$$

$$=\lim_{\epsilon\to 0}2\frac{t}{\tau}e^{\frac{2\epsilon t}{\tau}} \tag{124}$$

$$=2\frac{t}{\tau}. \tag{125}$$

which reveals the time dependent contribution of zero eigenvalues. Thus

$$e^{\mathbf{F}\frac{t}{\tau}}\mathbf{F}^{-1}e^{\mathbf{F}\frac{t}{\tau}}-\mathbf{F}^{-1}=\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{O}\mathbf{\Lambda}^{-1}\mathbf{O}^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T-\mathbf{O}\mathbf{\Lambda}^{-1}\mathbf{O}^T+4\frac{t}{\tau}\mathbf{M}\mathbf{M}^T. \tag{126}$$

We continue by substituting the above results into Fukumizu's equation

$$\begin{aligned}
\mathbf{Q}\mathbf{Q}^T(t)=&\left[\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T+2\mathbf{M}\mathbf{M}^T\right]\mathbf{Q}(0) && (127)\\
&\left[\mathbf{I}+\frac{1}{2}\mathbf{Q}(0)^T\left(\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{O}\mathbf{\Lambda}^{-1}\mathbf{O}^T\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T-\mathbf{O}\mathbf{\Lambda}^{-1}\mathbf{O}^T+4\frac{t}{\tau}\mathbf{M}\mathbf{M}^T\right)\mathbf{Q}(0)\right]^{-1}\\
&\mathbf{Q}(0)^T\left[\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T+2\mathbf{M}\mathbf{M}^T\right]\\
=&\left[\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T+2\mathbf{M}\mathbf{M}^T\right]\mathbf{Q}(0)\\
&\left[\mathbf{I}+\frac{1}{2}\mathbf{Q}(0)^T\left(\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{\Lambda}^{-1}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T-\mathbf{O}\mathbf{\Lambda}^{-1}\mathbf{O}^T+4\frac{t}{\tau}\mathbf{M}\mathbf{M}^T\right)\mathbf{Q}(0)\right]^{-1} && (128)\\
&\mathbf{Q}(0)^T\left[\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T+2\mathbf{M}\mathbf{M}^T\right]\\
=&\left[\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T+2\mathbf{M}\mathbf{M}^T\right]\mathbf{Q}(0)\\
&\left[\mathbf{I}+\frac{1}{2}\mathbf{Q}(0)^T\left(\mathbf{O}\left(e^{2\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{\Lambda}^{-1}-\mathbf{\Lambda}^{-1}\right)\mathbf{O}^T+4\frac{t}{\tau}\mathbf{M}\mathbf{M}^T\right)\mathbf{Q}(0)\right]^{-1} && (129)\\
&\mathbf{Q}(0)^T\left[\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T+2\mathbf{M}\mathbf{M}^T\right]\\
=&\left[\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T+2\mathbf{M}\mathbf{M}^T\right]\mathbf{Q}(0)\\
&\left[\mathbf{I}+\frac{1}{2}\mathbf{Q}(0)^T\left(\mathbf{O}\left(e^{2\mathbf{\Lambda}\frac{t}{\tau}}-\mathbf{I}\right)\mathbf{\Lambda}^{-1}\mathbf{O}^T+4\frac{t}{\tau}\mathbf{M}\mathbf{M}^T\right)\mathbf{Q}(0)\right]^{-1} && (130)\\
&\mathbf{Q}(0)^T\left[\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T+2\mathbf{M}\mathbf{M}^T\right].
\end{aligned}$$

Then, matrix multiplication on the left side of the equation yields

$$\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}=\frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}} & \tilde{\mathbf{V}}\\ \tilde{\mathbf{U}} & -\tilde{\mathbf{U}}\end{bmatrix}\begin{bmatrix}e^{\tilde{\mathbf{S}}\frac{t}{\tau}} & 0\\ 0 & e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\end{bmatrix} \tag{131}$$

$$=\frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}}e^{\tilde{\mathbf{S}}\frac{t}{\tau}} & \tilde{\mathbf{V}}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\\ \tilde{\mathbf{U}}e^{\tilde{\mathbf{S}}\frac{t}{\tau}} & -\tilde{\mathbf{U}}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\end{bmatrix} \tag{132}$$

and

$$\mathbf{O}^T\mathbf{Q}(0) = \frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}} & \tilde{\mathbf{V}}\\ \tilde{\mathbf{U}} & -\tilde{\mathbf{U}}\end{bmatrix}^T\begin{bmatrix}\mathbf{V}\sqrt{\mathbf{S}}\mathbf{R}^T\\ \mathbf{U}\sqrt{\mathbf{S}}\mathbf{R}^T\end{bmatrix} \tag{133}$$

$$= \frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}}^T\mathbf{V}\sqrt{\mathbf{S}}\mathbf{R}^T + \tilde{\mathbf{U}}^T\mathbf{U}\sqrt{\mathbf{S}}\mathbf{R}^T\\ \tilde{\mathbf{V}}^T\mathbf{V}\sqrt{\mathbf{S}}\mathbf{R}^T - \tilde{\mathbf{U}}^T\mathbf{U}\sqrt{\mathbf{S}}\mathbf{R}^T\end{bmatrix} \tag{134}$$

$$= \frac{1}{\sqrt{2}}\begin{bmatrix}\left(\tilde{\mathbf{V}}^T\mathbf{V} + \tilde{\mathbf{U}}^T\mathbf{U}\right)\sqrt{\mathbf{S}}\mathbf{R}^T\\ \left(\tilde{\mathbf{V}}^T\mathbf{V} - \tilde{\mathbf{U}}^T\mathbf{U}\right)\sqrt{\mathbf{S}}\mathbf{R}^T\end{bmatrix}, \tag{135}$$

such that

$$\mathbf{O}e^{\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{O}^T\mathbf{Q}(0) = \frac{1}{2}\begin{bmatrix}\tilde{\mathbf{V}}e^{\tilde{\mathbf{S}}\frac{t}{\tau}} & \tilde{\mathbf{V}}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\\ \tilde{\mathbf{U}}e^{\tilde{\mathbf{S}}\frac{t}{\tau}} & -\tilde{\mathbf{U}}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{V}}^T\mathbf{V}\sqrt{\mathbf{S}}\mathbf{R}^T + \tilde{\mathbf{U}}^T\mathbf{U}\sqrt{\mathbf{S}}\mathbf{R}^T\\ \tilde{\mathbf{V}}^T\mathbf{V}\sqrt{\mathbf{S}}\mathbf{R}^T - \tilde{\mathbf{U}}^T\mathbf{U}\sqrt{\mathbf{S}}\mathbf{R}^T\end{bmatrix} \tag{136}$$

$$= \frac{1}{2}\begin{bmatrix}\tilde{\mathbf{V}}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\left(\tilde{\mathbf{V}}^T\mathbf{V} + \tilde{\mathbf{U}}^T\mathbf{U}\right) + e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\left(\tilde{\mathbf{V}}^T\mathbf{V} - \tilde{\mathbf{U}}^T\mathbf{U}\right)\right)\sqrt{\mathbf{S}}\mathbf{R}^T\\ \tilde{\mathbf{U}}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\left(\tilde{\mathbf{V}}^T\mathbf{V} + \tilde{\mathbf{U}}^T\mathbf{U}\right) - e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\left(\tilde{\mathbf{V}}^T\mathbf{V} - \tilde{\mathbf{U}}^T\mathbf{U}\right)\right)\sqrt{\mathbf{S}}\mathbf{R}^T\end{bmatrix}. \tag{137}$$

We continue by calculating

$$4\mathbf{M}\mathbf{M}^T\mathbf{Q}(0) = 4\frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}}_\perp\\ \tilde{\mathbf{U}}_\perp\end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix}\tilde{\mathbf{V}}_\perp\\ \tilde{\mathbf{U}}_\perp\end{bmatrix}^T\begin{bmatrix}\mathbf{V}\sqrt{\mathbf{S}}\mathbf{R}^T\\ \mathbf{U}\sqrt{\mathbf{S}}\mathbf{R}^T\end{bmatrix} \tag{138}$$

$$= 2\begin{bmatrix}\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T & \tilde{\mathbf{V}}_\perp\tilde{\mathbf{U}}_\perp^T\\ \tilde{\mathbf{U}}_\perp\tilde{\mathbf{V}}_\perp^T & \tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\end{bmatrix}\begin{bmatrix}\mathbf{V}\sqrt{\mathbf{S}}\mathbf{R}^T\\ \mathbf{U}\sqrt{\mathbf{S}}\mathbf{R}^T\end{bmatrix} \tag{139}$$

$$= 2\begin{bmatrix}\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T & 0\\ 0 & \tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\end{bmatrix}\begin{bmatrix}\mathbf{V}\sqrt{\mathbf{S}}\mathbf{R}^T\\ \mathbf{U}\sqrt{\mathbf{S}}\mathbf{R}^T\end{bmatrix} \tag{140}$$

$$= 2\begin{bmatrix}\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V}\sqrt{\mathbf{S}}\mathbf{R}^T\\ \tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}\sqrt{\mathbf{S}}\mathbf{R}^T\end{bmatrix} \tag{141}$$

and

$$\frac{1}{2}\mathbf{Q}(0)^T 4\frac{t}{\tau}\mathbf{M}\mathbf{M}^T\mathbf{Q}(0) = \frac{t}{\tau}\begin{bmatrix}\mathbf{R}\sqrt{\mathbf{S}}\mathbf{V}^T & \mathbf{R}\sqrt{\mathbf{S}}\mathbf{U}^T\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V}\sqrt{\mathbf{S}}\mathbf{R}^T\\ \tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}\sqrt{\mathbf{S}}\mathbf{R}^T\end{bmatrix} \tag{142}$$

$$= \frac{t}{\tau}\left[\mathbf{R}\sqrt{\mathbf{S}}\left(\mathbf{V}^T\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V} + \mathbf{U}^T\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}\right)\sqrt{\mathbf{S}}\mathbf{R}^T\right] \tag{143}$$

Next, we define $\mathbf{B} = \mathbf{U}^T\tilde{\mathbf{U}} + \mathbf{V}^T\tilde{\mathbf{V}}$ and $\mathbf{C} = \mathbf{U}^T\tilde{\mathbf{U}} - \mathbf{V}^T\tilde{\mathbf{V}}$ and rewrite the inverse as

$$\left[\mathbf{I} + \frac{1}{2}\mathbf{Q}(0)^T\mathbf{O}\left(e^{2\mathbf{\Lambda}\frac{t}{\tau}} - \mathbf{I}\right)\mathbf{\Lambda}^{-1}\mathbf{O}^T\mathbf{Q}(0) + 2\frac{t}{\tau}\mathbf{Q}(0)^T\mathbf{M}\mathbf{M}^T\mathbf{Q}(0)\right]^{-1} \tag{144}$$

$$= \Bigg[\mathbf{I} + \frac{1}{4}\mathbf{R}\sqrt{\mathbf{S}}\Bigg(\begin{bmatrix}\mathbf{B} & -\mathbf{C}\end{bmatrix}\left(e^{2\mathbf{\Lambda}\frac{t}{\tau}} - \mathbf{I}\right)\mathbf{\Lambda}^{-1}\begin{bmatrix}\mathbf{B}^T\\ -\mathbf{C}^T\end{bmatrix} + 4\frac{t}{\tau}\left(\mathbf{V}^T\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V} + \mathbf{U}^T\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}\right)\Bigg)\sqrt{\mathbf{S}}\mathbf{R}^T\Bigg]^{-1}. \tag{145}$$

Working from the centre out, we have

$$\begin{bmatrix}\mathbf{B} & -\mathbf{C}\end{bmatrix}\mathbf{\Lambda}^{-1}\begin{bmatrix}\mathbf{B}^T\\ -\mathbf{C}^T\end{bmatrix} = \begin{bmatrix}\mathbf{B} & -\mathbf{C}\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{S}}^{-1} & 0\\ 0 & -\tilde{\mathbf{S}}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{B}^T\\ -\mathbf{C}^T\end{bmatrix} \tag{146}$$

$$= \begin{bmatrix}\mathbf{B} & -\mathbf{C}\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{S}}^{-1}\mathbf{B}^T\\ \tilde{\mathbf{S}}^{-1}\mathbf{C}^T\end{bmatrix} \tag{147}$$

$$= \mathbf{B}\tilde{\mathbf{S}}^{-1}\mathbf{B}^T - \mathbf{C}\tilde{\mathbf{S}}^{-1}\mathbf{C}^T \tag{148}$$

and

$$[\mathbf{B} \quad -\mathbf{C}]\, e^{2\mathbf{\Lambda}\frac{t}{\tau}}\mathbf{\Lambda}^{-1}\begin{bmatrix}\mathbf{B}^T\\-\mathbf{C}^T\end{bmatrix} = [\mathbf{B} \quad -\mathbf{C}]\begin{bmatrix}e^{2\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{S}}^{-1} & 0\\ 0 & -e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{S}}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{B}^T\\-\mathbf{C}^T\end{bmatrix} \tag{149}$$

$$= [\mathbf{B} \quad -\mathbf{C}]\begin{bmatrix}e^{2\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{S}}^{-1}\mathbf{B}^T\\ e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{S}}^{-1}\mathbf{C}^T\end{bmatrix} \tag{150}$$

$$= \mathbf{B}e^{2\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{S}}^{-1}\mathbf{B}^T - \mathbf{C}e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}}\tilde{\mathbf{S}}^{-1}\mathbf{C}^T. \tag{151}$$

Finally, using $AB^{-1} = (BA^{-1})^{-1}$ (and $A^{-1}B = (B^{-1}A)^{-1}$) to move terms into the inverse, we rewrite

$$\begin{aligned}
\mathbf{Q}\mathbf{Q}^T(t) = \frac{1}{2}&\begin{bmatrix}\tilde{\mathbf{V}}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^T - e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T\right) + 2\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V}\\ \tilde{\mathbf{U}}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^T + e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T\right) + 2\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}\end{bmatrix}\sqrt{\mathbf{S}}\mathbf{R}^T\\
&\Bigg[\mathbf{I} + \mathbf{R}\sqrt{\mathbf{S}}\Bigg(\frac{1}{4}\mathbf{B}\left(e^{2\tilde{\mathbf{S}}\frac{t}{\tau}} - \mathbf{I}\right)\tilde{\mathbf{S}}^{-1}\mathbf{B}^T - \frac{1}{4}\mathbf{C}\left(e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}} - \mathbf{I}\right)\tilde{\mathbf{S}}^{-1}\mathbf{C}^T\\
&+ \frac{t}{\tau}\left(\mathbf{V}^T\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V} + \mathbf{U}^T\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}\right)\Bigg)\sqrt{\mathbf{S}}\mathbf{R}^T\Bigg]^{-1}\\
&\frac{1}{2}\begin{bmatrix}\tilde{\mathbf{V}}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^T - e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T\right) + 2\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V}\\ \tilde{\mathbf{U}}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^T + e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T\right) + 2\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}\end{bmatrix}\sqrt{\mathbf{S}}\mathbf{R}^T\Bigg]^T\\
= \frac{1}{2}&\begin{bmatrix}\tilde{\mathbf{V}}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^T - e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T\right) + 2\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V}\\ \tilde{\mathbf{U}}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^T + e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T\right) + 2\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}\end{bmatrix}\\
&\Bigg[\mathbf{S}^{-1} + \frac{1}{4}\mathbf{B}\left(e^{2\tilde{\mathbf{S}}\frac{t}{\tau}} - \mathbf{I}\right)\tilde{\mathbf{S}}^{-1}\mathbf{B}^T - \frac{1}{4}\mathbf{C}\left(e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}} - \mathbf{I}\right)\tilde{\mathbf{S}}^{-1}\mathbf{C}^T\\
&+ \frac{t}{\tau}\left(\mathbf{V}^T\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V} + \mathbf{U}^T\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}\right)\Bigg]^{-1}\\
&\frac{1}{2}\begin{bmatrix}\tilde{\mathbf{V}}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^T - e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T\right) + 2\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V}\\ \tilde{\mathbf{U}}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^T + e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T\right) + 2\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}\end{bmatrix}^T\\
= &\begin{bmatrix}\tilde{\mathbf{V}}\left(\mathbf{I} - e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\right) + 2\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V}(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\\ \tilde{\mathbf{U}}\left(\mathbf{I} + e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\right) + 2\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\end{bmatrix}\\
&\Bigg[4e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^{-1}\mathbf{S}^{-1}(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}} + \left(\mathbf{I} - e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}}\right)\tilde{\mathbf{S}}^{-1}\\
&- e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^{-1}\mathbf{C}\left(e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}} - \mathbf{I}\right)\tilde{\mathbf{S}}^{-1}\mathbf{C}^T(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\\
&+ 4\frac{t}{\tau}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^{-1}\left(\mathbf{V}^T\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V} + \mathbf{U}^T\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}\right)(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\Bigg]^{-1}\\
&\begin{bmatrix}\tilde{\mathbf{V}}\left(\mathbf{I} - e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\right) + 2\tilde{\mathbf{V}}_\perp\tilde{\mathbf{V}}_\perp^T\mathbf{V}B^{-T}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\\ \tilde{\mathbf{U}}\left(\mathbf{I} + e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\right) + 2\tilde{\mathbf{U}}_\perp\tilde{\mathbf{U}}_\perp^T\mathbf{U}B^{-T}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\end{bmatrix}^T.
\end{aligned} \tag{152--154}$$

C.3 Proof of Theorem 3.2: Limiting behaviour

As training time increases, all terms including a matrix exponential with negative exponent in Equation 11 vanish to zero, as $\tilde{\mathbf{S}}$ is a diagonal matrix with entries larger zero

$$\lim_{t\to\infty} e^{-\tilde{\mathbf{S}}\frac{t}{\tau}} = \mathbf{0}. \tag{155}$$

Therefore, in the temporal limit, eq. 11 reduces to

$$\lim_{t\to\infty} \mathbf{Q}\mathbf{Q}^T(t) = \lim_{t\to\infty} \begin{bmatrix} \mathbf{W}_1^T\mathbf{W}_1(t) & \mathbf{W}_1^T\mathbf{W}_2^T(t) \\ \mathbf{W}_2\mathbf{W}_1(t) & \mathbf{W}_2^T\mathbf{W}_2(t) \end{bmatrix} \tag{156}$$

$$= \begin{bmatrix} \tilde{\mathbf{V}} \\ \tilde{\mathbf{U}} \end{bmatrix} \left[\tilde{\mathbf{S}}^{-1}\right]^{-1} \begin{bmatrix} \tilde{\mathbf{V}}^T & \tilde{\mathbf{U}}^T \end{bmatrix} \tag{157}$$

$$= \begin{bmatrix} \tilde{\mathbf{V}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T & \tilde{\mathbf{V}}\tilde{\mathbf{S}}\tilde{\mathbf{U}}^T \\ \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T & \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{U}}^T \end{bmatrix}. \tag{158}$$

□

C.4 Dynamics of $\mathbf{Q}(t)$

The solution for the weights $\mathbf{W}_1(t)$ and $\mathbf{W}_2(t)$ can be derived up to a time varying orthogonal transformation as demonstrated by Yan et al. [61].

Under the assumptions of whitened inputs 2.2, zero-balanced weights 2.3, full rank 2.4, and equal input-output dimension, the temporal dynamics of $\mathbf{Q}(t)$ is given as

$$\mathbf{Q}(t) = e^{\mathbf{F}\frac{t}{\tau}}\mathbf{Q}(0)\left[\mathbf{I} + \frac{1}{2}\mathbf{Q}(0)^T\left(e^{\mathbf{F}\frac{t}{\tau}}\mathbf{F}^{-1}e^{\mathbf{F}\frac{t}{\tau}} - \mathbf{F}^{-1}\right)\mathbf{Q}(0)\right]^{-\frac{1}{2}}\mathbf{D}(t). \tag{159}$$

where $\mathbf{D}(t)$ is an orthogonal matrix of size $N_h \times N_h$. From this definition, computing $\mathbf{Q}(t)\mathbf{Q}(t)^T$, we recover equation 47.

Equation 159 shows that the individual weight matrices are not directly described by parts of the $\mathbf{Q}(t)\mathbf{Q}(t)^T$ solution. Instead, they are fixed only up to a time-dependent orthogonal transformation. To verify this, we numerically compute $\mathbf{D}(t)$ as $\mathbf{D}(t) = \mathbf{q}(t)^+\mathbf{Q}_{\text{sim}}(t)$ where $\mathbf{Q}_{\text{sim}}(t)$ denotes weights obtained from numerical simulations of gradent descent, $+$ denotes the pseudoinverse ($\mathbf{q}^+(t) = (\mathbf{q}^T(t)\mathbf{q}(t))^{-1}\mathbf{q}(\mathbf{t})^T$ where $\mathbf{q}(t)$ is rectangular) and

$$\mathbf{q}(t) = e^{\mathbf{F}\frac{t}{\tau}}\mathbf{Q}(0)\left[\mathbf{I} + \frac{1}{2}\mathbf{Q}(0)^T\left(e^{\mathbf{F}\frac{t}{\tau}}\mathbf{F}^{-1}e^{\mathbf{F}\frac{t}{\tau}} - \mathbf{F}^{-1}\right)\mathbf{Q}(0)\right]^{-\frac{1}{2}}. \tag{160}$$

We numerically show in Fig. 7D right panel that $\mathbf{D}(t)$ generally changes over time. Letting $\mathbf{Q}_d(t)$ denote the estimated $\mathbf{Q}(t)$ using the numerically recovered $\mathbf{D}(t)$, Fig. 7D left and centre panels show that both the dynamics of $\mathbf{Q}_d(t)$ and $\mathbf{Q}_d(t)\mathbf{Q}_d(t)^T$ match the temporal dynamics of the simulation. The small derivation between the simulation and the analytical solution for later time points, is due to the imprecision of the pseudoinverse.

In Fig. 7C, we report the implementation of equation 160. As expected, the analytical solution does not match the numerical temporal dynamics. However,the solution for $\mathbf{q}(t)\mathbf{q}(t)^T$ recovers the correct dynamics.

D Appendix: Rich and lazy learning regimes and generalisation

Under the assumptions of Theorem 3.1, the network function acquires a rich task-specific internal representation at convergence, that is $\mathbf{W}_1^T\mathbf{W}_1 = \tilde{\mathbf{V}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T$ and $\mathbf{W}_2\mathbf{W}_2^T = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{U}}^T$. Therefore, there exist initial states with large zero-balanced weights that lead to rich solutions.

We more quantitatively capture this phenomena in Fig. 8. We define the error on the internal representation as Fig. 3 $||\mathbf{W}_1^T\mathbf{W}_1 - \tilde{\mathbf{V}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T||_F^2$ and $||\mathbf{W}_2\mathbf{W}_2^T - \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{U}}^T|_F^2$ for $\mathbf{W}_1$ and $\mathbf{W}_2$ respectively. Effectively, we measure the richness of the representation and in turn it's generalisation ability. In Fig. 8, the error remains zero for increasing gain for any network initialised with zero-balanced weights. In other words, the representation at convergences is rich. In contrast, for random initialisation the error increase consequently with increasing gain. As the network is moving away from the small random weight initialisation, the network converges to lazier representation.

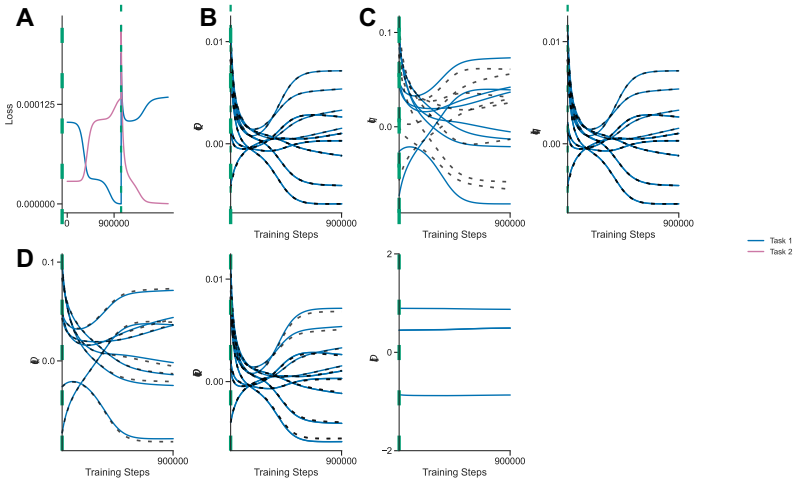

Figure 7: **A:** Loss under gradient descent learning two random input-output correlation task with learning rate $\eta = 0,001$ up to precision $1e-7$. The green dotted line marks the time at which the target is switched from task 1 to task 2. **B:** Numerical (coloured line) and analytical (black dotted line) temporal dynamics of $\mathbf{Q}\mathbf{Q}^T(t)$ as given by eq. 161. **C:** Numerical (coloured line) and analytical (black dotted line) temporal dynamics of $\mathbf{q}(t)$ and $\mathbf{q}(t)\mathbf{q}(t)^T$ 160 **D:** Temporal dynamics of $\mathbf{D}(t)$. Numerical (coloured line) and analytical (black dotted line) temporal dynamics of $\mathbf{Q}_d(t)\mathbf{Q}_d(t)^T$ and $\mathbf{Q}_d(t)$ as given by equation 159 where $\mathbf{D}$ was computed numerically.

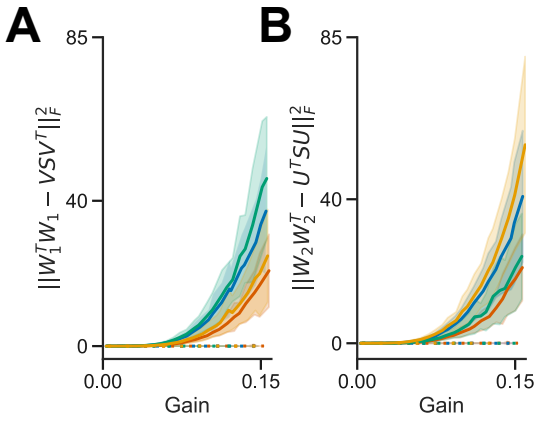

Figure 8: **A.B** Mean and standard deviation on the error on the internal representation error defined as in section D for the learning the living kingdom task (Fig. 6A), a random 7×7 matrix (blue), a random 5×7 matrix (yellow), a 7×5 matrix (green), a 8×8 matrix (red). All the task ran were ran with learning rate $\eta = 0.001$ enforcing initial zero-balanced weights 2.3 (dotted line) and breaking the assumption of zero-balanced initial weights 2.3 (line). $N_h = 10$ for all networks.

E Appendix: Decoupling dynamics

E.1 Proof for Theorem 5.1

Let the input and output dimension of a two-layer linear network (eq. 1) be equal, i.e., $N_i = N_o$, then eq. 11 simplifies to

$$\begin{aligned}\mathbf{Q}\mathbf{Q}^T(t) = &\begin{bmatrix}\tilde{\mathbf{V}}\left(\mathbf{I} - e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\right)\\ \tilde{\mathbf{U}}\left(\mathbf{I} + e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\right)\end{bmatrix}\\ &\Big[4e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^{-1}\mathbf{S}^{-1}(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}} + \left(\mathbf{I} - e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}}\right)\tilde{\mathbf{S}}^{-1}\\ &- e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^{-1}\mathbf{C}\left(e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}} - \mathbf{I}\right)\tilde{\mathbf{S}}^{-1}\mathbf{C}^T(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\Big]^{-1}\\ &\begin{bmatrix}\tilde{\mathbf{V}}\left(\mathbf{I} - e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\right)\\ \tilde{\mathbf{U}}\left(\mathbf{I} + e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T(\mathbf{B}^T)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\right)\end{bmatrix}^T.\end{aligned} \tag{161}$$

Further, let the singular value decomposition of the input-output correlation of the task be

$$\mathrm{SVD}(\tilde{\mathbf{\Sigma}}^{yx}) = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T \tag{162}$$

and suppose that the initial state of the network can be written in the form

$$\mathrm{SVD}(\mathbf{W}_2(0)\mathbf{W}_1(0)) = \mathbf{U}\mathbf{S}\mathbf{V}^T = \tilde{\mathbf{U}}\mathbf{A}(0)^T\mathbf{A}(0)\tilde{\mathbf{V}}^T. \tag{163}$$

First, we note that the initial weights in this setting are not independent of the structure of the target task. In particular,

$$\mathbf{U}\sqrt{\mathbf{S}} = \tilde{\mathbf{U}}\mathbf{A}(0)^T \tag{164}$$

$$\Leftrightarrow \tilde{\mathbf{U}}^T\mathbf{U}\sqrt{\mathbf{S}} = \mathbf{A}(0)^T \tag{165}$$

$$\Leftrightarrow \sqrt{\mathbf{S}}\mathbf{U}^T\tilde{\mathbf{U}} = \mathbf{A}(0) \tag{166}$$

$$\tag{167}$$

and

$$\sqrt{\mathbf{S}}\mathbf{V}^T = \mathbf{A}(0)\tilde{\mathbf{V}}^T \tag{168}$$

$$\Leftrightarrow \sqrt{\mathbf{S}}\mathbf{V}^T\tilde{\mathbf{V}} = \mathbf{A}(0) \tag{169}$$

and therefore

$$\sqrt{\mathbf{S}}\mathbf{U}^T\tilde{\mathbf{U}} = \sqrt{\mathbf{S}}\mathbf{V}^T\tilde{\mathbf{V}} \tag{170}$$

$$\Leftrightarrow \mathbf{U}\mathbf{V}^T = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^T. \tag{171}$$

This further simplifies the equation, as

$$\mathbf{U}\sqrt{\mathbf{S}} = \tilde{\mathbf{U}}\mathbf{A}(0)^T \tag{172}$$

$$\Leftrightarrow \mathbf{U} = \tilde{\mathbf{U}}\mathbf{A}(0)^T\sqrt{\mathbf{S}}^{-1} \tag{173}$$

and

$$\sqrt{\mathbf{S}}\mathbf{V}^T = \mathbf{A}(0)\tilde{\mathbf{V}}^T \tag{174}$$

$$\Leftrightarrow \mathbf{V}^T = \sqrt{\mathbf{S}}^{-1}\mathbf{A}(0)\tilde{\mathbf{V}}^T \tag{175}$$

$$\Leftrightarrow \mathbf{V} = \tilde{\mathbf{V}}\mathbf{A}(0)^T\sqrt{\mathbf{S}}^{-1}, \tag{176}$$

then recollecting the definition of $\mathbf{B}$ and $\mathbf{C}$ we get

$$\mathbf{B}^T = \tilde{\mathbf{U}}^T\mathbf{U} + \tilde{\mathbf{V}}^T\mathbf{V} \tag{177}$$

$$= \tilde{\mathbf{U}}^T\tilde{\mathbf{U}}\mathbf{A}(0)^T\sqrt{S}^{-1} + \tilde{\mathbf{V}}^T\tilde{\mathbf{V}}\mathbf{A}(0)^T\sqrt{S}^{-1} \tag{178}$$

$$= \left(\tilde{\mathbf{U}}^T\tilde{\mathbf{U}} + \tilde{\mathbf{V}}^T\tilde{\mathbf{V}}\right)\mathbf{A}(0)^T\sqrt{S}^{-1} \tag{179}$$

$$= 2\mathbf{A}(0)^T\sqrt{S}^{-1} \tag{180}$$

and

$$\mathbf{C}^T = \tilde{\mathbf{U}}^T\mathbf{U} - \tilde{\mathbf{V}}^T\mathbf{V} \tag{181}$$

$$= \left(\tilde{\mathbf{U}}^T\tilde{\mathbf{U}} - \tilde{\mathbf{V}}^T\tilde{\mathbf{V}}\right)\mathbf{A}(0)^T\sqrt{S}^{-1} \tag{182}$$

$$= 0. \tag{183}$$

Substituting the new values of $\mathbf{B}$ and $\mathbf{C}$ into Equation 161 then yields

$$\mathbf{Q}\mathbf{Q}^T(t) = \begin{bmatrix}\tilde{\mathbf{V}}\\ \tilde{\mathbf{U}}\end{bmatrix}\left[4e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\frac{1}{4}\mathbf{A}(0)^{-1}\sqrt{\mathbf{S}}\mathbf{S}^{-1}\sqrt{\mathbf{S}}\mathbf{A}(0)^{-T}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}} + \left(\mathbf{I} - e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}}\right)\tilde{\mathbf{S}}^{-1}\right]^{-1}\begin{bmatrix}\tilde{\mathbf{V}}\\ \tilde{\mathbf{U}}\end{bmatrix}^T \tag{184}$$

$$= \begin{bmatrix}\tilde{\mathbf{V}}\\ \tilde{\mathbf{U}}\end{bmatrix}\left[e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\left(\mathbf{A}(0)^T\mathbf{A}(0)\right)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}} + \left(\mathbf{I} - e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}}\right)\tilde{\mathbf{S}}^{-1}\right]^{-1}\begin{bmatrix}\tilde{\mathbf{V}}\\ \tilde{\mathbf{U}}\end{bmatrix}^T. \tag{185}$$

Finally, we note that the dynamics can thus be written as

$$\mathbf{Q}\mathbf{Q}^T(t) = \begin{bmatrix}\tilde{\mathbf{V}}\mathbf{A}^T\mathbf{A}(t)\tilde{\mathbf{V}}^T & \tilde{\mathbf{V}}\mathbf{A}^T\mathbf{A}(t)\tilde{\mathbf{U}}^T\\ \tilde{\mathbf{U}}\mathbf{A}^T\mathbf{A}(t)\tilde{\mathbf{V}}^T & \tilde{\mathbf{U}}\mathbf{A}^T\mathbf{A}(t)\tilde{\mathbf{U}}^T\end{bmatrix} \tag{186}$$

where

$$\mathbf{A}^T\mathbf{A}(t) = \left[e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\left(\mathbf{A}(0)^T\mathbf{A}(0)\right)^{-1}e^{-\tilde{\mathbf{S}}\frac{t}{\tau}} + \left(\mathbf{I} - e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}}\right)\tilde{\mathbf{S}}^{-1}\right]^{-1}. \tag{187}$$

□

E.2 Solution for 2×2 dynamics

We consider small networks with input and output dimension $N_i = 2$ and $N_o = 2$. In this setting, the structure of the weight initialisation and task are encoded in the matrices

$$\mathbf{A}(0)^T\mathbf{A}(0) = \begin{bmatrix}a_1(0) & b(0)\\ b(0) & a_2(0)\end{bmatrix} \quad \text{and} \quad \tilde{\mathbf{S}} = \begin{bmatrix}s_1 & 0\\ 0 & s_2\end{bmatrix}, \tag{188}$$

where the parameters $a_1(0)$ and $a_2(0)$ represent coupling within a singular mode, and $b(0)$ represents counterproductive cross-coupling between different singular modes.

From Equation 13, we have

$$\mathbf{A}^T\mathbf{A}(t) = \left[\begin{bmatrix}e^{\frac{-s_1t}{\tau}} & 0\\ 0 & e^{\frac{-s_2t}{\tau}}\end{bmatrix}\begin{bmatrix}a_1(0) & b(0)\\ b(0) & a_2(0)\end{bmatrix}^{-1}\begin{bmatrix}e^{\frac{-s_1t}{\tau}} & 0\\ 0 & e^{\frac{-s_2t}{\tau}}\end{bmatrix}\right. \tag{189}$$

$$\left. + \left[\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} - \begin{bmatrix}e^{\frac{-2s_1t}{\tau}} & 0\\ 0 & e^{\frac{-2s_2t}{\tau}}\end{bmatrix}\right]\begin{bmatrix}s_1 & 0\\ 0 & s_2\end{bmatrix}^{-1}\right]^{-1} \tag{190}$$

$$= \left[\frac{1}{a_1(0)a_2(0) - b(0)^2}\begin{bmatrix}e^{\frac{-s_1t}{\tau}} & 0\\ 0 & e^{\frac{-s_2t}{\tau}}\end{bmatrix}\begin{bmatrix}a_2(0) & -b(0)\\ -b(0) & a_1(0)\end{bmatrix}\begin{bmatrix}e^{\frac{-s_1t}{\tau}} & 0\\ 0 & e^{\frac{-s_2t}{\tau}}\end{bmatrix}\right.$$

$$\left. + \left[\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} - \begin{bmatrix}e^{\frac{-2s_1t}{\tau}} & 0\\ 0 & e^{\frac{-2s_2t}{\tau}}\end{bmatrix}\right]\begin{bmatrix}\frac{1}{s_1} & 0\\ 0 & \frac{1}{s_2}\end{bmatrix}\right]^{-1},$$

where we use

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \tag{191}$$

We continue with

$$\mathbf{A}^T\mathbf{A}(t) = \Bigg[\frac{1}{a_1(0)a_2(0)-b(0)^2}\begin{bmatrix} e^{\frac{-s_1t}{\tau}} & 0 \\ 0 & e^{\frac{-s_2t}{\tau}} \end{bmatrix}\begin{bmatrix} a_2(0) & -b(0) \\ -b(0) & a_1(0) \end{bmatrix}\begin{bmatrix} e^{\frac{-s_1t}{\tau}} & 0 \\ 0 & e^{\frac{-s_2t}{\tau}} \end{bmatrix} \tag{192}$$

$$+\begin{bmatrix} \frac{1}{s_1} & 0 \\ 0 & \frac{1}{s_2} \end{bmatrix} - \begin{bmatrix} \frac{1}{s_1}e^{\frac{-2s_1t}{\tau}} & 0 \\ 0 & \frac{1}{s_2}e^{\frac{-2s_2t}{\tau}} \end{bmatrix}\Bigg]^{-1} \tag{193}$$

$$= \Bigg[\frac{1}{a_1(0)a_2(0)-b(0)^2}\begin{bmatrix} e^{\frac{-2s_1t}{\tau}}a_2(0) & -e^{\frac{-s_1t}{\tau}}b(0)e^{\frac{-s_2t}{\tau}} \\ -e^{\frac{-s_2t}{\tau}}b(0)e^{\frac{-s_1t}{\tau}} & e^{\frac{-2s_2t}{\tau}}a_1(0) \end{bmatrix}$$

$$+\begin{bmatrix} \frac{1}{s_1} & 0 \\ 0 & \frac{1}{s_2} \end{bmatrix} - \begin{bmatrix} \frac{1}{s_1}e^{\frac{-2s_1t}{\tau}} & 0 \\ 0 & \frac{1}{s_2}e^{\frac{-2s_2t}{\tau}} \end{bmatrix}\Bigg]^{-1}$$

$$= \begin{bmatrix} \frac{e^{\frac{-2s_1t}{\tau}}a_2(0)}{a_1(0)a_2(0)-b(0)^2} + \frac{1}{s_1} - \frac{1}{s_1}e^{\frac{-2s_1t}{\tau}} & -\frac{e^{\frac{-s_1t}{\tau}}b(0)e^{\frac{-s_2t}{\tau}}}{a_1(0)a_2(0)-b(0)^2} \\ -\frac{e^{\frac{-s_2t}{\tau}}b(0)e^{\frac{-s_1t}{\tau}}}{a_1(0)a_2(0)-b(0)^2} & \frac{e^{\frac{-2s_2t}{\tau}}a_1(0)}{a_1(0)a_2(0)-b(0)^2} + \frac{1}{s_2} - \frac{1}{s_2}e^{\frac{-2s_2t}{\tau}} \end{bmatrix}^{-1} \tag{194}$$

.

We use equation 191 and simplify the denominator

$$\mathbf{A}^T\mathbf{A}(t) =$$

$$\frac{1}{\Big(\frac{e^{\frac{-2s_2t}{\tau}}a_1(0)}{a_1(0)a_2(0)-b(0)^2} + \frac{1}{s_2} - \frac{1}{s_2}e^{\frac{-2s_2t}{\tau}}\Big)\Big(\frac{e^{\frac{-2s_1t}{\tau}}a_2(0)}{a_1(0)a_2(0)-b(0)^2} + \frac{1}{s_1} - \frac{1}{s_1}e^{\frac{-2s_1t}{\tau}}\Big) - \Big(-\frac{e^{\frac{-s_2t}{\tau}}b(0)e^{\frac{-s_1t}{\tau}}}{a_1(0)a_2(0)-b(0)^2}\Big)^2} \tag{195}$$

$$\times\begin{bmatrix} \frac{e^{\frac{-2s_2t}{\tau}}a_1(0)}{a_1(0)a_2(0)-b(0)^2} + \frac{1}{s_2} - \frac{1}{s_2}e^{\frac{-2s_2t}{\tau}} & \frac{e^{\frac{-s_1t}{\tau}}b(0)e^{\frac{-s_2t}{\tau}}}{a_1(0)a_2(0)-b(0)^2} \\ \frac{e^{\frac{-s_2t}{\tau}}b(0)e^{\frac{-s_1t}{\tau}}}{a_1(0)a_2(0)-b(0)^2} & \frac{e^{\frac{-2s_1t}{\tau}}a_2(0)}{a_1(0)a_2(0)-b(0)^2} + \frac{1}{s_1} - \frac{1}{s_1}e^{\frac{-2s_1t}{\tau}} \end{bmatrix}.$$

The diagonal element $a_1(t)$ is given as

$$a_1(t) = \frac{\frac{e^{\frac{-2s_2t}{\tau}}a_1(0)}{a_1(0)a_2(0)-b(0)^2} + \frac{1}{s_2} - \frac{1}{s_2}e^{\frac{-2s_2t}{\tau}}}{\Big(\frac{e^{\frac{-2s_2t}{\tau}}a_1(0)}{a_1(0)a_2(0)-b(0)^2} + \frac{1}{s_2} - \frac{1}{s_2}e^{\frac{-2s_2t}{\tau}}\Big)\Big(\frac{e^{\frac{-2s_1t}{\tau}}a_2(0)}{a_1(0)a_2(0)-b(0)^2} + \frac{1}{s_1} - \frac{1}{s_1}e^{\frac{-2s_1t}{\tau}}\Big) - \Big(-\frac{e^{\frac{-s_2t}{\tau}}b(0)e^{\frac{-s_1t}{\tau}}}{a_1(0)a_2(0)-b(0)^2}\Big)^2}, \tag{196}$$

and interchanging subscripts 1 and 2 yields $a_2(t)$. As a check on this result, by setting $b(0) = 0$ we recover the expression

$$a_1(t) = \frac{a_1(0)}{e^{\frac{-2s_1t}{\tau}} + \frac{a_1(0)}{s_1}\Big(1 - e^{\frac{-2s_1t}{\tau}}\Big)}, \tag{197}$$

from Saxe et al. [25].

We further simplify the denominator to

$$\mathbf{A}^T\mathbf{A}(t) =$$

$$\frac{1}{\frac{1}{a_1(0)a_2(0)-b(0)^2}\big(e^{\frac{-2(s_1+s_2)t}{\tau}}(1 - \frac{a_1(0)}{s_1} - \frac{a_2(0)}{s_2}) + e^{\frac{-2s_2t}{\tau}}\frac{a_1(0)}{s_1} + e^{\frac{-2s_1t}{\tau}}\frac{a_2(0)}{s_2}\big) + \frac{1}{s_2s_1}} \tag{198}$$

$$\times\begin{bmatrix} \frac{e^{\frac{-2s_2t}{\tau}}a_1(0)}{a_1(0)a_2(0)-b(0)^2} + \frac{1}{s_2} - \frac{1}{s_2}e^{\frac{-2s_2t}{\tau}} & \frac{e^{\frac{-s_1t}{\tau}}b(0)e^{\frac{-s_2t}{\tau}}}{a_1(0)a_2(0)-b(0)^2} \\ \frac{e^{\frac{-s_2t}{\tau}}b(0)e^{\frac{-s_1t}{\tau}}}{a_1(0)a_2(0)-b(0)^2} & \frac{e^{\frac{-2s_1t}{\tau}}a_2(0)}{a_1(0)a_2(0)-b(0)^2} + \frac{1}{s_1} - \frac{1}{s_1}e^{\frac{-2s_1t}{\tau}} \end{bmatrix}$$

E.3 Off-Diagonal decoupling dynamics

We track the decoupling by considering the dynamics of the off-diagonal element $b(t)$.

$$b(t) = \frac{\frac{e^{\frac{-s_2 t}{\tau}} b(0) e^{\frac{-s_1 t}{\tau}}}{a_1(0)a_2(0)-b(0)^2}}{\frac{1}{a_1(0)a_2(0)-b(0)^2}\left(e^{\frac{-2(s_1+s_2)t}{\tau}}(1-\frac{a_1(0)}{s_1}-\frac{a_2(0)}{s_2})+e^{\frac{-2s_2 t}{\tau}}\frac{a_1(0)}{s_1}+e^{\frac{-2s_1 t}{\tau}}\frac{a_2(0)}{s_2}\right)+\frac{1}{s_2 s_1}}. \tag{199}$$

As t tends to infinity $\lim_{t\to\infty} b(t) = 0$ the off-diagonal element shrinks to zero.

We can further simplify the off-diagonal to

$$b(t) = \frac{b(0)}{e^{\frac{-(s_1+s_2)t}{\tau}}(1-\frac{a_1(0)}{s_1}-\frac{a_2(0)}{s_2})+e^{\frac{(s_1-s_2)t}{\tau}}\frac{a_1(0)}{s_1}+e^{\frac{(s_2-s_1)t}{\tau}}\frac{a_2(0)}{s_2}+\frac{a_1(0)a_2(0)-b(0)^2}{s_2 s_1}}. \tag{200}$$

Equation 200 can exhibit non-monotonic trajectories with transient peaks as shown in Fig. 4. The qualitative observations for the 2×2 network hold for larger target matrices as shown in Fig. 9. For large initialisation, the dynamics are exponential. At intermediate and small initialisation, the maximum of the off-diagonal is reached before the singular mode is fully learned. In the small initialisation scheme, the peak is of negligible size. The respective target matrix for Panel A-D, B-E and C-F in Fig. 9 are

$$\text{dense } \begin{bmatrix} 5 & 6 & 3 & 0 & 1 \\ 4, & 1 & 0 & 1 & 2 \\ 3 & 0 & 2 & 4 & 0 \\ 3 & 4 & 0 & 3 & 2 \\ 2 & 0 & 1 & 3 & 4 \end{bmatrix}, \text{ diagonal } \begin{bmatrix} 5 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix} \text{ and equal diagonal } \begin{bmatrix} 5 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix}.$$

We characterise these dynamics considering the case where $s_1 = s_2 = s$ for the two-by-two solution (i.e. equal diagonal target $\mathbf{y}$) for which we can compute the time of the peak. In this particular case, we can further simplify the off-diagonal to

$$b(t) = \frac{b(0)}{e^{\frac{-2(s)t}{\tau}}(1-\frac{a_1(0)+a_2(0)}{s})+\frac{a_1(0)+a_2(0)}{s}+\frac{a_1(0)a_2(0)-b(0)^2}{s^2}}. \tag{201}$$

We find the time of the maximum of the off-diagonal elements to be $t_{peak} = \frac{\tau}{4s}\ln\frac{s(s-a_1(0)-a_2(0))}{a_1(0)a_2(0)-b(0)^2}$.

The presence of a peak in the off-diagonal values, indicates the decoupling, but as shown in Fig. 4D-F, the peak size is negligible in comparison to the size of the on-diagonal values for small initial weights. This difference is reminiscent of the silent alignment effect described by [26]. We further note, that the time scale of decoupling is on the same order as the one reported for the silent alignment effect $t_{sa} = \frac{1}{s}$.

E.4 On-diagonal dynamics and the effect of initialisation variance

In this section we revisit the impact of initialisation scale for the on-diagonal dynamics. We now start with

$$a_1(t) = \frac{\frac{e^{\frac{-2s_2 t}{\tau}} a_1(0)}{a_1(0)a_2(0)-b(0)^2}+\frac{1}{s_2}-\frac{1}{s_2}e^{\frac{-2s_2 t}{\tau}}}{\frac{1}{a_1(0)a_2(0)-b(0)^2}\left(e^{\frac{-2(s_1+s_2)t}{\tau}}(1-\frac{a_1(0)}{s_1}-\frac{a_2(0)}{s_2})+e^{\frac{-2s_2 t}{\tau}}\frac{a_1(0)}{s_1}+e^{\frac{-2s_1 t}{\tau}}\frac{a_2(0)}{s_2}\right)+\frac{1}{s_2 s_1}}. \tag{202}$$

The diagonal elements simplify in the cases where $s_1 = s_2 = s$ (i.e. target $\mathbf{Y}$ is diagonal),

$$a_1(t) = \frac{\frac{e^{\frac{-2st}{\tau}} a_1(0)}{a_1(0)a_2(0)-b(0)^2}+\frac{1}{s}-\frac{1}{s}e^{\frac{-2st}{\tau}}}{\frac{1}{a_1(0)a_2(0)-b(0)^2}\left(e^{\frac{-4st}{\tau}}(1-\frac{a_1(0)}{s}-\frac{a_2(0)}{s})+e^{\frac{-2st}{\tau}}\frac{a_1(0)}{s}+e^{\frac{-2st}{\tau}}\frac{a_2(0)}{s}\right)+\frac{1}{s^2}}. \tag{203}$$

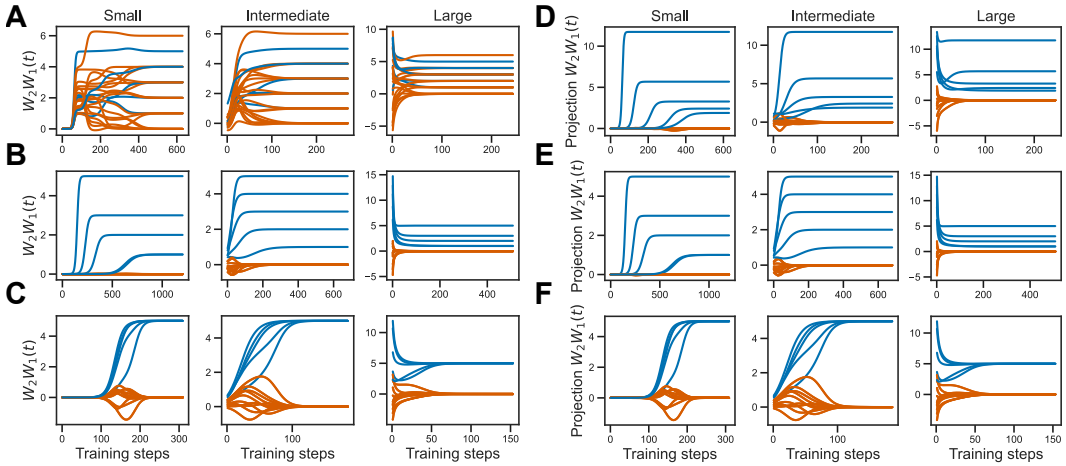

Figure 9: **A-C** Network function dynamics (Diagonal elements: blue, Off-diagonal elements: red) learning with learning rate $\eta = 0.01$ on the target 5×5 diagonal matrices shown in Equation 201. The network was initialised as defined in Section E with Small ($\sigma = 1e{-}6$), Intermediate ($\sigma = 0.1$) and Large ($\sigma = 2$) variance, and hidden layer size $N_h = 10$. **A**, Dense. **B**, Diagonal. **C**, Equal diagonal. **D-F**. Corresponding numerical temporal dynamics of the projection of the network function on- and off-diagonal elements into the singular-basis of the initialisation. Equivalently, the temporal dynamics of the elements of $\mathbf{A}\mathbf{A}^T$ bottom left quadrant. **D**, Dense. **E**, Diagonal. **F**, Equal diagonal.

We consider when $|a_1(0)|, |a_2(0)|, |b(0)| \ll 1$, and recover a sigmoidal trajectory,

$$a_1(t) = \frac{s a_1(0)}{e^{\frac{-2st}{\tau}} \left[s - a_1(0) - a_2(0)\right] + a_1(0) + a_2(0)}. \tag{204}$$

We can compute the time at which $a_1(t)$ rises to half its asymptotic value to be

$$t_{\text{half}} = \frac{\tau}{2s} \log\left(\frac{s - a_1(0) - a_2(0)}{a_1(0) - a_2(0)}\right). \tag{205}$$

For $|a_1(0)|, |a_2(0)|, |b(0)| \gg 0$ the dynamics of the on-diagonal element a_1 is close to exponential.

The observation for 2×2 network hold for larger target matrices as shown in Fig. 9. For large variance initialisations, the dynamics are exponential. At intermediate variance initialisations, we observe more complex behaviour. While at small variance initialisations, the on-diagonal element describes a sigmoidal trajectory.

F Appendix: Continual Learning

We consider the case of training a two-layer deep linear network on a sequence of tasks $\mathcal{T}_a$, $\mathcal{T}_b$, $\mathcal{T}_c$, ... with corresponding correlation functions $\mathcal{T}_a = \tilde{\mathbf{\Sigma}}^{yx}_a$, $\mathcal{T}_b = \tilde{\mathbf{\Sigma}}^{yx}_b$.... Then, the full batch loss of the i-th task at any point in training time is

$$\mathcal{L}_i = \frac{1}{2P} ||\mathbf{W}_2\mathbf{W}_1\mathbf{X}_i - \mathbf{Y}_i||^2_F. \tag{206}$$

From Theorem 3.2 it follows that after training the network to convergence on task $\mathcal{T}_j$, the network function is $\mathbf{W}_2\mathbf{W}_1 = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T = \tilde{\mathbf{\Sigma}}^{yx}_j$. Further, using the assumption of whitened inputs 2.2 and the identities $||A||^2_F = tr(AA^T)$ and $tr(A) + tr(B) = tr(A + B)$, the full batch loss of the i-th task is

then

$$\mathcal{L}_i(\mathcal{T}_j) = \frac{1}{2P} \left|\left| \tilde{\mathbf{\Sigma}}_j^{yx} \mathbf{X}_i - \mathbf{Y}_i \right|\right|_F^2 \tag{207}$$

$$= \frac{1}{2P} \operatorname{Tr}\left((\tilde{\mathbf{\Sigma}}_j^{yx} \mathbf{X}_i - \mathbf{Y}_i|)(\tilde{\mathbf{\Sigma}}_j^{yx} \mathbf{X}_i - \mathbf{Y}_i|)^T \right) \tag{208}$$

$$= \frac{1}{2P} \operatorname{Tr}\left(\tilde{\mathbf{\Sigma}}_j^{yx} \mathbf{X}_i \mathbf{X}_i^T \tilde{\mathbf{\Sigma}}_j^{yx^T} \right) - \frac{1}{P} \operatorname{Tr}\left(\tilde{\mathbf{\Sigma}}_j^{yx} \mathbf{X}_i \mathbf{Y}_i^T \right) + \frac{1}{2P} \operatorname{Tr}\left(\mathbf{Y}_i \mathbf{Y}_i^T \right) \tag{209}$$

$$= \frac{1}{2} \operatorname{Tr}\left(\tilde{\mathbf{\Sigma}}_j^{yx} \tilde{\mathbf{\Sigma}}_j^{yx^T} \right) - \operatorname{Tr}\left(\tilde{\mathbf{\Sigma}}_j^{yx} \tilde{\mathbf{\Sigma}}_i^{yx^T} \right) + \frac{1}{2} \operatorname{Tr}\left(\tilde{\mathbf{\Sigma}}_i^{yy} \right) \tag{210}$$

$$= \frac{1}{2} \operatorname{Tr}\left(\left(\tilde{\mathbf{\Sigma}}_j^{yx} - \tilde{\mathbf{\Sigma}}_i^{yx} \right) \left(\tilde{\mathbf{\Sigma}}_j^{yx} - \tilde{\mathbf{\Sigma}}_i^{yx} \right)^T - \tilde{\mathbf{\Sigma}}_i^{yx} \tilde{\mathbf{\Sigma}}_i^{yx^T} \right) + \frac{1}{2} \left(\tilde{\mathbf{\Sigma}}_i^{yy} \right) \tag{211}$$

$$= \frac{1}{2} \left|\left| \tilde{\mathbf{\Sigma}}_j^{yx} - \tilde{\mathbf{\Sigma}}_i^{yx} \right|\right|_F^2 \underbrace{- \frac{1}{2} \operatorname{Tr}\left(\tilde{\mathbf{\Sigma}}_i^{yx} \tilde{\mathbf{\Sigma}}_i^{yx^T} \right) + \frac{1}{2} \left(\tilde{\mathbf{\Sigma}}_i^{yy} \right)}_{c}. \tag{212}$$

Therefore, the amount of forgetting $\mathcal{F}$ on task $\mathcal{T}_i$ when training on task $\mathcal{T}_k$ after having trained the network on task $\mathcal{T}_j$, i.e. the relative change of loss, is fully determined by the similarity structure of the tasks

$$\mathcal{F}_i(\mathcal{T}_j, \mathcal{T}_k) = \mathcal{L}_i(\mathcal{T}_k) - \mathcal{L}_i(\mathcal{T}_j) \tag{213}$$

$$= \frac{1}{2} \left|\left| \tilde{\mathbf{\Sigma}}_k^{yx} - \tilde{\mathbf{\Sigma}}_i^{yx} \right|\right|_F^2 + c - \frac{1}{2} \left|\left| \tilde{\mathbf{\Sigma}}_j^{yx} - \tilde{\mathbf{\Sigma}}_i^{yx} \right|\right|_F^2 - c \tag{214}$$

$$= \frac{1}{2} \left(\left|\left| \tilde{\mathbf{\Sigma}}_k^{yx} - \tilde{\mathbf{\Sigma}}_i^{yx} \right|\right|_F^2 - \left|\left| \tilde{\mathbf{\Sigma}}_j^{yx} - \tilde{\mathbf{\Sigma}}_i^{yx} \right|\right|_F^2 \right). \tag{215}$$

G Appendix: Revising structured knowledge

G.1 Reversal learning dynamics

In the following, we assume that the input dimension is equal to the output dimension. Further, we denote the i-th column of the left and right singular vectors as $\mathbf{u}_i$, $\tilde{\mathbf{u}}_i$ and $\mathbf{v}_i$, $\tilde{\mathbf{v}}_i$ respectively.

Reversal learning occurs when the task and the initial network function share the same left and right singular vectors, i.e., $\mathbf{U} = \tilde{\mathbf{U}}$ and $\mathbf{V} = \tilde{\mathbf{V}}$, except for one or multiple columns of the left singular vectors, for which the direction is reversed:

$$-\mathbf{u}_i = \tilde{\mathbf{u}}_i. \tag{216}$$

We note that, if there is any reversal in the right singular vectors $-\mathbf{v}_i = \tilde{\mathbf{v}}_i$, this can be written as a reversal in the left singular vectors, as the signs of the right and left singular vectors are interchangeable. In the reversal learning setting, both $\mathbf{B} = \mathbf{U}^T \tilde{\mathbf{U}} + \mathbf{V}^T \tilde{\mathbf{V}}$ and $\mathbf{C} = \mathbf{U}^T \tilde{\mathbf{U}} - \mathbf{V}^T \tilde{\mathbf{V}}$ are diagonal matrices. The diagonal entries of $\mathbf{C}$ are zero if the singular vectors are aligned and 2 if they are reversed. Similarly, diagonal entries of $\mathbf{B}$ are 2 if the singular vectors are aligned and zero if they are reversed. Therefore, in the case of reversal learning, $\mathbf{B}$ is a diagonal matrix with 0 values and thus is not invertible. As a consequence, the learning dynamics cannot be described by Equation 11. However, as $\mathbf{B}$ and $\mathbf{C}$ are diagonal matrices, the learning dynamics simplify. Let $\mathbf{b}_i$, $\mathbf{c}_i$, $\mathbf{s}_i$ and $\tilde{\mathbf{s}}_i$ denote the i-th diagonal entry of $\mathbf{B}$, $\mathbf{C}$, $\mathbf{S}$ and $\tilde{\mathbf{S}}$ respectively, then the network dynamics

can be rewritten as

$$\begin{aligned}\mathbf{W}_2\mathbf{W}_1(t) = &\frac{1}{2}\tilde{\mathbf{U}}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B}^T + e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}^T\right) \\ &\left[\mathbf{S}^{-1} + \frac{1}{4}\mathbf{B}\left(e^{2\tilde{\mathbf{S}}\frac{t}{\tau}} - \mathbf{I}\right)\tilde{\mathbf{S}}^{-1}\mathbf{B}^T - \frac{1}{4}\mathbf{C}\left(e^{-2\tilde{\mathbf{S}}\frac{t}{\tau}} - \mathbf{I}\right)\tilde{\mathbf{S}}^{-1}\mathbf{C}^T\right]^{-1} \qquad (217) \\ &\frac{1}{2}\left(e^{\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{B} - e^{-\tilde{\mathbf{S}}\frac{t}{\tau}}\mathbf{C}\right)\tilde{\mathbf{V}}^T \\ = &\sum_{i=1}^{N_i}\frac{\mathbf{b}_i^2 e^{2\tilde{\mathbf{s}}_i\frac{t}{\tau}} - \mathbf{c}_i^2 e^{-2\tilde{\mathbf{s}}_i\frac{t}{\tau}}}{4\mathbf{s}_i^{-1} + \mathbf{b}_i^2 e^{2\tilde{\mathbf{s}}_i\frac{t}{\tau}}\tilde{\mathbf{s}}_i^{-1} - \mathbf{b}_i^2\tilde{\mathbf{s}}_i^{-1} - \mathbf{c}_i^2 e^{-2\tilde{\mathbf{s}}_i\frac{t}{\tau}}\tilde{\mathbf{s}}_i^{-1} + \mathbf{c}_i^2\tilde{\mathbf{s}}_i^{-1}}\,\tilde{\mathbf{u}}_i\tilde{\mathbf{v}}_i^T \qquad (218) \\ = &\sum_{i=1}^{N_i}\frac{\mathbf{s}_i\mathbf{b}_i^2\tilde{\mathbf{s}}_i - \mathbf{s}_i\mathbf{c}_i^2\tilde{\mathbf{s}}_i e^{-4\tilde{\mathbf{s}}_i\frac{t}{\tau}}}{4\tilde{\mathbf{s}}_i e^{-2\tilde{\mathbf{s}}_i\frac{t}{\tau}} + \mathbf{s}_i\mathbf{b}_i^2\left(1 - e^{-2\tilde{\mathbf{s}}_i\frac{t}{\tau}}\right) + \mathbf{s}_i\mathbf{c}_i^2\left(e^{-2\tilde{\mathbf{s}}_i\frac{t}{\tau}} - e^{-4\tilde{\mathbf{s}}_i\frac{t}{\tau}}\right)}\tilde{\mathbf{u}}_i\tilde{\mathbf{v}}_i^T. \qquad (219)\end{aligned}$$

It follows, that in the reversal learning case, i.e. $\mathbf{b} = 0$, for each reversed singular vector, the dynamics vanish to zero

$$\lim_{t\to\infty}\frac{-\mathbf{s}_i\mathbf{c}_i^2\tilde{\mathbf{s}}_i e^{-4\tilde{\mathbf{s}}_i\frac{t}{\tau}}}{4\tilde{\mathbf{s}}_i e^{-2\tilde{\mathbf{s}}_i\frac{t}{\tau}} + \mathbf{s}_i\mathbf{c}_i^2\left(e^{-2\tilde{\mathbf{s}}_i\frac{t}{\tau}} - e^{-4\tilde{\mathbf{s}}_i\frac{t}{\tau}}\right)}\tilde{\mathbf{u}}_i\tilde{\mathbf{v}}_i^T = 0. \qquad (220)$$

Analytically, the learning dynamics are initialised and remain on the separatrix of a saddle point, until the corresponding singular value of the network function has vanished and remains zero, corresponding to convergence to the saddle point. When simulated numerically, the learning dynamics escape the saddle points due to imprecision of floating point arithmetic. However, numerical optimisation still suffers from catastrophic slowing [60], as escaping the saddle point takes time (Fig. 6A). In contrast, in the case of aligned singular vectors ($\mathbf{c} = 0$), we recover the equation for the temporal dynamics as described in Saxe et al. [17]. Training succeeds, as the singular value of the network function converges to its target value

$$\begin{aligned}\lim_{t\to\infty}\sum_{i=1}^{N_i}\frac{\mathbf{s}_i\mathbf{b}_i^2\tilde{\mathbf{s}}_i}{4\tilde{\mathbf{s}}_i e^{-2\tilde{\mathbf{s}}_i\frac{t}{\tau}} + \mathbf{s}_i\mathbf{b}_i^2\left(1 - e^{-2\tilde{\mathbf{s}}_i\frac{t}{\tau}}\right)}\tilde{\mathbf{u}}_i\tilde{\mathbf{v}}_i^T &= \frac{\mathbf{s}_i\mathbf{b}_i^2\tilde{\mathbf{s}}_i}{\mathbf{s}_i\mathbf{b}_i^2}\tilde{\mathbf{u}}_i\tilde{\mathbf{v}}_i^T \qquad (221) \\ &= \tilde{\mathbf{s}}_i\tilde{\mathbf{u}}_i\tilde{\mathbf{v}}_i^T. \qquad (222)\end{aligned}$$

In summary, in the case of aligned singular vectors, the learning dynamics can be described by the convergence of singular values. However in the case of reversal learning, analytically, training does not succeed. In simulations, the learning dynamics escape the saddle point due to numerical imprecision, but the learning dynamics are catastrophically slowed in the vicinity of the saddle point.

G.2 Exact learning dynamics in shallow networks

To provide a point of comparison to our deep linear network results, here we derive a solution for the temporal dynamics of reversal learning in a shallow network.

The network's weights are optimised using full batch gradient descent with learning rate η (or equivalently time constant $\tau = 1/\eta$) on the mean squared error loss given in Equation 2, yielding the first task dynamics

$$\tau\frac{d}{dt}\mathbf{W} = \tilde{\mathbf{\Sigma}}^{yx} - \mathbf{W}\tilde{\mathbf{\Sigma}}^{xx}, \qquad (223)$$

where $\tilde{\mathbf{\Sigma}}^{xx}$ and $\tilde{\mathbf{\Sigma}}^{yx}$ is the input and input-output correlation matrices of the dataset. We define

$$\mathrm{SVD}(\mathbf{W}(0)) = \mathbf{U}\mathbf{S}\mathbf{V}^T \quad \text{and} \quad \mathrm{SVD}(\tilde{\mathbf{\Sigma}}^{yx}) = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T. \qquad (224)$$

motivating the change of variable $\mathbf{W} = \mathbf{U}\overline{\mathbf{W}}\mathbf{V}^T$. We project the weight into the basis of the initialisation

$$\tau \frac{d}{dt}\mathbf{U}\overline{\mathbf{W}}\mathbf{V}^T = \tilde{\mathbf{\Sigma}}^{yx} - \mathbf{U}\overline{\mathbf{W}}\mathbf{V}^T\tilde{\mathbf{\Sigma}}^{xx} \tag{225}$$

$$\tau \frac{d}{dt}\mathbf{U}\overline{\mathbf{W}}\mathbf{V}^T = \mathbf{U}\mathbf{U}^T\tilde{\mathbf{\Sigma}}^{yx}\mathbf{V}\mathbf{V}^T - \mathbf{U}\overline{\mathbf{W}}\mathbf{V}^T\tilde{\mathbf{\Sigma}}^{xx} \tag{226}$$

$$\tau \frac{d}{dt}\overline{\mathbf{W}} = \mathbf{U}^T\tilde{\mathbf{\Sigma}}^{yx}\mathbf{V} - \overline{\mathbf{W}}\tilde{\mathbf{\Sigma}}^{xx}. \tag{227}$$

Under the assumption of whitened inputs 2.2, the dynamics yields

$$\tau \frac{d}{dt}\overline{\mathbf{W}} = \mathbf{U}^T\tilde{\mathbf{\Sigma}}^{yx}\mathbf{V} - \overline{\mathbf{W}}. \tag{228}$$

Defining $\overline{\mathbf{W}}_{ii} = b_i$ the diagonal element of the matrix, encoding the strength of the mode i transmitted by the input-to-output weight. Similarly, we write $(\mathbf{U}^T\tilde{\mathbf{\Sigma}}^{yx}\mathbf{V})_{ii} = k_i$. Assuming decoupled initial conditions, we obtain the scalar dynamics

$$\tau \frac{d}{dt}b_i = k_i - b_i \tag{229}$$

with solution

$$b_i = k_i(1 - e^{\frac{-t}{\tau}}) + b_i^0 e^{\frac{-t}{\tau}}. \tag{230}$$

Reverting the change of variable, the weight trajectory yields

$$\mathbf{W} = \mathbf{U}\mathbf{B}(t)\mathbf{V}^T. \tag{231}$$

This solution is very similar to the one proposed by Saxe et al. [25]. However, the key here is that k_i can have negative values. k_i is negative whenever a vector is in the opposite direction to the initialisation (as in the reversal learning setting). We show in Fig. 6 that the analytical solution derived above matches the numerical temporal dynamics. From Equation 230, we note that the shallow network cannot display catastrophic slowing.

H Simulations

In the following, we describe the details of the simulation studies. Generally, N_i, N_h and N_o denote the dimension of the input, hidden layer and output (target) respectively. The number of training samples is N and the learning rate is denoted by $\eta = 1/\tau$.

H.1 Zero-balanced weight initialisation

The initial network weights are zero-balanced 2.3 when they satisfy

$$\mathbf{W}_1(0)\mathbf{W}_1(0)^T = \mathbf{W}_2(0)^T\mathbf{W}_2(0). \tag{232}$$

In practice, we use Algorithm 1 to initialise the network weights, where α is a scaling factor which is used to control the variance of the weights, i.e., to vary between small and large weight initialisations.

H.2 Tasks

In the following, we describe the different tasks that are used throughout the simulation studies.

H.2.1 Random regression task

In a random regression task the inputs $\mathbf{X} \in \mathbb{R}^{N_i,N}$ are sampled from a random normal distribution $\mathbf{X} \sim \mathcal{N}(\mu = 0, \sigma = 1)$. The input data $\mathbf{X}$ is then whitened, such that $1/N\mathbf{X}\mathbf{X}^T = \mathbf{I}$. The target values $\mathbf{Y} \in \mathbb{R}^{N_o,N}$ are also sampled from a random normal distribution, however, with variance adjusted to the number of output nodes $\mathbf{Y} \sim \mathcal{N}(\mu = 0, \alpha = 1/\sqrt{N_o})$. Thus, network inputs and target values are uncorrelated Gaussian noise and therefore, a linear solution does not always exist.

Algorithm 1 Zero-balanced weight initialisation

Require: N_i, N_h, N_o, σ
  $\mathbf{W}_1 \sim \mathcal{N}(\mu = 0, \sigma) \in \mathbb{R}^{N_h \times N_i}$
  $\mathbf{W}_2 \sim \mathcal{N}(\mu = 0, \sigma) \in \mathbb{R}^{N_o \times N_h}$
  $\mathbf{U}, \mathbf{S}, \mathbf{V} \leftarrow \text{SVD}(\mathbf{W}_2\mathbf{W}_1)$
  $\mathbf{S} \leftarrow \sqrt{\mathbf{S}}$
  $\mathbf{R} \sim \mathcal{N}(\mu = 0, \sigma = 1) \in \mathbb{R}^{N_h \times N_h}$
  $\mathbf{R}, \_, \_ \leftarrow \text{SVD}(\mathbf{R})$
  if $N_i \neq N_o$ **then**
    $N_s \leftarrow N_i$ **if** $N_i < N_o$ **else** N_o
    $\mathbf{S}_1 \leftarrow \begin{bmatrix} \mathbf{S} \\ \mathbf{0}_{N_h - N_s \times N_s} \end{bmatrix}$
    $\mathbf{S}_2 \leftarrow [\mathbf{S} \quad \mathbf{0}_{N_s \times N_h - N_s}]$
    $\mathbf{W}_1 \leftarrow \mathbf{R}\mathbf{S}_1\mathbf{V}^T$
    $\mathbf{W}_2 \leftarrow \mathbf{U}\mathbf{S}_2\mathbf{R}^T$
  else
    $\mathbf{W}_1 \leftarrow \mathbf{R}\mathbf{S}\mathbf{V}^T$
    $\mathbf{W}_2 \leftarrow \mathbf{U}\mathbf{S}\mathbf{R}^T$
  end if
  return $\mathbf{W}_1\mathbf{W}_2$

H.2.2 Teacher-student task

In order to guarantee that a linear solution exists, we use the teacher-student setup. First, inputs $\mathbf{X}$ are sampled as in the random regression task. Then, target values $\mathbf{Y}$ are generated by sampling a pair of random zero-balanced weights $\mathbf{W}_1 \in \mathbb{R}^{N_h \times N_i}$ and $\mathbf{W}_2 \in \mathbb{R}^{N_o \times N_h}$ and then calculating $\mathbf{Y} = \mathbf{W}_2\mathbf{W}_1\mathbf{X}$. Like this, it is ensured that a linear solution exists. The variance of the output is varied by changing the variation within the zero-balanced weights σ.

H.2.3 Semantic hierarchy

Input items in the semantic hierarchy task are encoded as one-hot vectors, i.e. $\mathbf{X} = \mathbf{I}$. The corresponding target vectors y_i encoded the position in the hierarchical tree. Where a 1 encoded being a left child of a node, a -1 encoded being a right child of a node and a 0 encoded that the item is not a child of that node. For example, the blue fish is a blue fish, it is a left child of the root node, a left child of the animal node, not part of the plant branch, a right child of the fish node, and not part of the bird, algae or flower branch, leading to the label $[1, 1, 1, 0, -1, 0, 0, 0]$. The labels for all objects in the semantic tree as depicted in Figure 3A is then

$$\mathbf{Y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}. \tag{233}$$

The singular value decomposition for the corresponding correlation matrix $\tilde{\mathbf{\Sigma}}^{yx}$ are not unique. The first two, the third and the fourth and the last four singular values are identical. In order to match the numerical and analytical solution, this permutation invariance is removed by adding a small constant perturbation to each column $\mathbf{y}_i, i \in 1, ..., N$ of the labels

$$\mathbf{y}_i = \mathbf{y}_i * (1 + \frac{0.1}{i}), \tag{234}$$

leading to almost but not exactly identical singular values.

H.2.4 Colour hierarchy

Following the same procedure as described for the semantic hierarchy, the labels for the colour hierarchy as depicted in Figure 6C are then

$$\mathbf{Y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 \\ 0 & -1 & 1 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \end{bmatrix}. \tag{235}$$

H.3 Figure 1

Figure 1 panels B-D show three simulations from varying initial weights on the same teacher-student task. The task was created with $\sigma = 0.35$. Farther, $N_i = 5$, $N_h = 10$, $N_o = 2$ and $N = 10$. The learning rate was $\eta = 0.1$ and the initial network weights were sampled with $\sigma = 0.01$, $\sigma = 0.25$ and $\sigma = 0.25$ in panels B, C and D respectively.

H.4 Figure 2

Figure 2 panels A and B show a simulation on the same teacher-student task ($\sigma = 0.25$), once from small initial weights ($\sigma = 0.01$) and once from large initial weights ($\sigma = 0.15$). Dimensions were $N_i = 4$, $N_h = 5$, $N_o = 3$ and $N = 10$ and the learning rate was $\eta = 0.05$. Panel C was generated by running 50 simulations, each with a different initial random seed. For each of the simulations, dimensions were sampled randomly, such that $N_i \in [2, 50]$, $N_o \in [2, 50]$, $N_h = [\min(N_i, N_o), 50]$ and $N \in [2\max(N_i, N_h, N_o), 3\max(N_i, N_h, N_o)]$. Then, a random regression task was generated. Subsequently, a linear network was initialised with $\sigma \sim \mathcal{U}[{}^{0.01}/_{\sqrt{\max(N_i,\, N_o,\, N_h)}},\, {}^{0.5}/_{\sqrt{\max(N_i,\, N_o,\, N_h)}}]$. The network was then trained until convergence on the same task from the same initial weights for seven different learning rates $\eta \in \{0.05, 0.0232, 0.0107, 0.005, 0.0023, 0.0011, 0.0005\}$.

H.5 Figure 3

Panels C-F in Figure 3 were generated by training a linear network with $N_i = 8$, $N_h = 14$, $N_o = 8$ on the $N = 8$ items of the semantic hierarchy task. The learning rate was $\eta = 0.05$ and the initial weights in panels C, D, and E were sampled from a normal distribution with $\sigma = 0.0001$ and $\sigma = 0.42$ and zero-balanced weights with $\sigma = 0.44$ respectively.

H.6 Figure 4

Figure. 4 panel A was generated by training a linear network with $N_i = 5$, $N_h = 10$, $N_o = 5$ on the target $\mathbf{Y}$ as shown in Equation 201 (equal diagonal). The network was initialised with $\sigma = 0.1$. The learning rate was $\eta = 0.01$.

Figure. 4 panel D, E and F was generated by training a linear network with $N_i = 2$, $N_h = 10$, $N_o = 2$ on the target $\mathbf{Y}$ as shown in Figure. 4C and input $\mathbf{X} = bfi$. The network was initialised with small $\sigma = 0.00001$, intermediate $\sigma = 0.3$ and large $\sigma = 2$ synaptic weights. The learning rate was $\eta = 0.0001$.

H.7 Figure 5

Figure. 5 panel A was generated by training a linear network with $N_i = 5$, $N_h = 10$, $N_o = 6$ subsequently on four different random regression tasks with $N = 25$. The learning rate was $\eta = 0.05$ and the initial weights were small ($\sigma = 0.0001$).

Panels B and C were generated by running 50 simulations on two subsequent random regression tasks, each with a different initial random seed. The simulation was repeated three times, the first time with a linear, the second time with a $\tanh$ and the last time with a ReLU activation function in the hidden layer. Dimension were randomly sampled such that $N_i \in [2, 30]$, $N_o \in [2, 30]$,

$N_h = [\min(N_i, N_o), 30]$ and $N = 100$. The standard deviation of the initial weight was chosen such that $\sigma = {}^{0.5}/\sqrt{0.5(N_i + N_h)}$. The learning rate was $\eta = 0.075$.

For panel D and E the same simulation was repeated for three times, the first time with a linear, the second time with a $\tanh$ and the last time with a ReLU activation function. Each time, five random regression tasks with dimensions $N_i = 15$, $N_h = 18$, $N_o = 21$ and $N = 50$ were generated. Then a network with initial weight scale $\alpha = 0.025$ was sequentially trained with learning rate $\eta = 0.1$ on the five random regression tasks.

H.8 Figure 6

Figure 6 panel A was generated by training a linear network with $N_i = 4$, $N_h = 6$, $N_o = 4$ on a reversal learning task (see Section G.1), which was derived from a random regression task. The learning rate was $\eta = 0.05$ and initial weights had a standard deviation of $\sigma = 0.25$. Panel B was generated by training a shallow linear network (see Section G.2) on the same reversal learning task, with identical hyperparameters as in panel A.

For the top and bottom rows of panels E-F a linear network with $N_i = 8$, $N_h = 14$, $N_o = 8$ was trained on the semantic hierarchy task, followed by training the network on the adapted semantic hierarchy as depicted in Figure 6 C top, which is a reversal learning task and the colour hierarchy respectively. The learning rate was $\eta = 0.05$ and σ was set to 0.001 and 0.35 respectively.