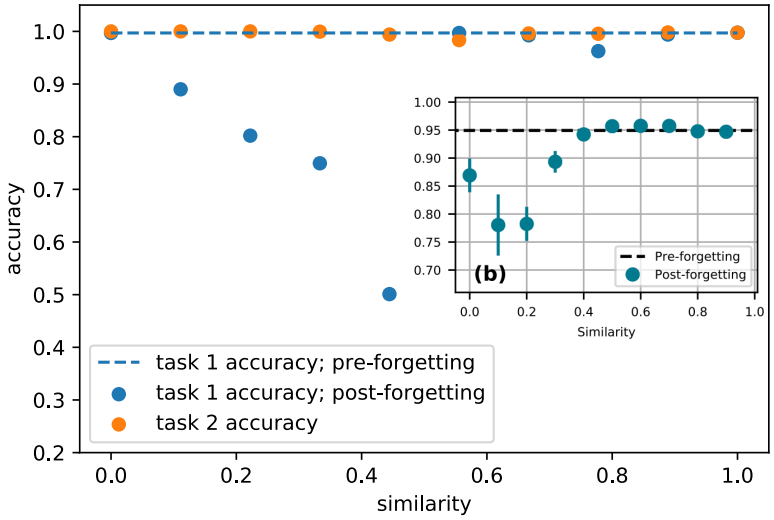

Figure 6. **Task similarity in deep vs. shallow networks** We plot the accuracy of a two-layer ReLU network with 8 neurons trained on two tasks. The first task is discriminating T-shirts from high-heels on Fashion MNIST (task 1). The second task is a linear interpolation in both inputs and labels between task 1 and long-sleeve shirts vs trainers. In the inset, we reproduce Fig. 5b of Ramasesh et al. (2020) when training various deep networks on two tasks obtained by linearly interpolation of CIFAR10 images. *Parameters of our experiment*: learning rate 0.01, $D = 784$.

A. Reproducing the results of Ramasesh et al. with two-layer neural networks

We report in Fig. 6 a reproduction of an experiment showing that the two-layer networks trained on FashionMNIST (Xiao et al., 2017) reproduce a key observation of (Ramasesh et al., 2020) made for VGG, ResNet and DenseNet on CIFAR10: intermediate task similarity leads to worst forgetting. To that end, we trained a two-layer ReLU network with 8 neurons to discriminate T-shirts from high-heels on Fashion MNIST (task 1). The second task was a linear interpolation in both inputs and labels between task 1 and long-sleeve shirts vs trainers. We see that at intermediate task similarity, or halfway along the linear interpolation between the two datasets, forgetting of the first task is the worst. This is the same behaviour Ramasesh et al. (2020) found consistently for VGG, ResNet and DenseNet when linearly interpolating CIFAR10 images (we reproduce their Fig. 5b in the inset). Hence the toy model studied here reproduces this behaviour of more realistic setups.

B. Order Parameters

The full set of order parameters for the two-teacher student-teacher networks in the large input limit is given by:

$$\text{Student-Student Overlap, } \mathbf{Q}: q_{kl} \equiv \langle \lambda_k \lambda_l \rangle = \frac{1}{N} \mathbf{w}_k \mathbf{w}_l; \tag{B.1}$$

$$\text{Teacher}^\dagger\text{-Teacher}^\dagger\text{Overlap, } \mathbf{T}: t_{nm} \equiv \langle \rho_m \rho_n \rangle = \frac{1}{N} \mathbf{w}_m^\dagger \mathbf{w}_n^\dagger; \tag{B.2}$$

$$\text{Student-Teacher}^\dagger\text{Overlap, } \mathbf{R}: r_{km} \equiv \langle \lambda_k \rho_m \rangle = \frac{1}{N} \mathbf{w}_k \mathbf{w}_m^\dagger; \tag{B.3}$$

$$\text{Teacher}^\ddagger\text{-Teacher}^\ddagger\text{Overlap, } \mathbf{S}: s_{pq} \equiv \langle \eta_p \eta_q \rangle = \frac{1}{N} \mathbf{w}_p^\ddagger \mathbf{w}_q^\ddagger; \tag{B.4}$$

$$\text{Student-Teacher}^\ddagger\text{Overlap, } \mathbf{U}: u_{kp} \equiv \langle \lambda_k \eta_p \rangle = \frac{1}{N} \mathbf{w}_k \mathbf{w}_p^\ddagger; \tag{B.5}$$

$$\text{Teacher}^\dagger\text{-Teacher}^\ddagger\text{Overlap, } \mathbf{V}: v_{mp} \equiv \langle \rho_m \eta_p \rangle = \frac{1}{N} \mathbf{w}_m^\dagger \mathbf{w}_p^\ddagger. \tag{B.6}$$

C. ODE Derivation

This section presents the derivation of the ODE formulation of the generalisation error for the student-multi-teacher continual learning framework.

C.1. Generalisation Error in terms of Order Parameters

Our aim is to formulate the generalisation error in terms of the macroscopic order parameters. Let us begin by multiplying out Eq. 2,

$$\epsilon_g^\dagger = \frac{1}{2}\left\langle \left[\sum_{i,k} h_i^\dagger h_k^\dagger g(\lambda_i) g(\lambda_k) + \sum_{m,n} v_m^\dagger v_n^\dagger g(\rho_m) g(\rho_n) - 2 \sum_{i,n} h_i^\dagger v_n^\dagger g(\lambda_i) g(\rho_n) \right] \right\rangle . \tag{C.1}$$

and similarly for the second student. These generalisation errors involve averages of local fields, which can be computed as integrals over a joint multivariate Gaussian probability distribution, all of the form

$$\mathcal{P}(\beta, \gamma) = \frac{1}{\sqrt{(2\pi)^{F+H} |\tilde{\mathbf{C}}|}} \exp\left\{ -\frac{1}{2} (\beta, \gamma)^T \tilde{\mathbf{C}}^{-1} (\beta, \gamma) \right\}, \tag{C.2}$$

where β and γ are local fields with number of units F and H respectively, and $\tilde{\mathbf{C}}$ is a covariance matrix suitably projected down from

$$\mathbf{C} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} & \mathbf{U} \\ \mathbf{R}^T & \mathbf{T} & \mathbf{V} \\ \mathbf{U}^T & \mathbf{V}^T & \mathbf{S} \end{pmatrix}.$$

We define

$$I_2(f, h) \equiv \langle g(\beta) g(\gamma) \rangle, \tag{C.3}$$

where f, h are the indices corresponding to the units of the local fields β and γ. This allows us to write the generalisation errors as

$$\epsilon_g^\dagger = \frac{1}{2} \sum_{i,k} h_i^\dagger h_k^\dagger I_2(i,k) + \frac{1}{2} \sum_{n,m} v_n^\dagger v_m^\dagger I_2(n,m) - \sum_{i,n} h_i^\dagger v_n^\dagger I_2(i,n) \tag{C.4}$$

$$\epsilon_g^\ddagger = \frac{1}{2} \sum_{i,k} h_i^\ddagger h_k^\ddagger I_2(i,k) + \frac{1}{2} \sum_{p,q} v_p^\ddagger v_q^\ddagger I_2(p,q) - \sum_{i,p} h_i^\ddagger v_p^\ddagger I_2(i,p). \tag{C.5}$$

C.1.1. Sigmoidal Activation

For the scaled error activation function, $g(x) = \text{erf}(x/\sqrt{2})$, there is an analytic expression for the I_2 integral purely in terms of the order parameters (Saad & Solla, 1995a):

$$I_2(i,k) = \frac{1}{\pi} \arcsin \frac{q_{ik}}{\sqrt{(1+q_{ii})(1+q_{kk})}}. \tag{C.6}$$

In turn, we can similarly write the generalisation errors in terms of the order parameters only:

$$\epsilon_g^\dagger = \frac{1}{\pi} \sum_{i,k} h_i^\dagger h_k^\dagger \arcsin \frac{q_{ik}}{\sqrt{(1+q_{ii})(1+q_{kk})}} + \frac{1}{\pi} \sum_{n,m} v_n^\dagger v_m^\dagger \arcsin \frac{t_{nm}}{\sqrt{(1+t_{nn})(1+t_{mm})}} + \frac{2}{\pi} \sum_{i,n} h_i^\dagger v_n^\dagger \arcsin \frac{r_{in}}{\sqrt{(1+q_{ii})(1+t_{nn})}} \tag{C.7}$$

$$\epsilon_g^\ddagger = \frac{1}{\pi} \sum_{i,k} h_i^\ddagger h_k^\ddagger \arcsin \frac{q_{ik}}{\sqrt{(1+q_{ii})(1+q_{kk})}} + \frac{1}{\pi} \sum_{p,q} v_p^\ddagger v_q^\ddagger \arcsin \frac{s_{pq}}{\sqrt{(1+s_{pp})(1+s_{qq})}} + \frac{2}{\pi} \sum_{i,p} h_i^\ddagger v_p^\ddagger \arcsin \frac{u_{ip}}{\sqrt{(1+q_{ii})(1+s_{pp})}}. \tag{C.8}$$

C.2. Order Parameter Evolution (Training on †)

Having arrived at expressions for the generalisation error of both teachers in terms of the order parameters, we want to determine equations of motion for these order parameters from the weight update equations (Eq. 5a & Eq. 5b). Trivially, the order parameters associated with the two teachers, $\mathbf{T}$ and $\mathbf{S}$ are constant over time, as are the head weights of the teachers, $\mathbf{v}^\dagger, \mathbf{v}^\ddagger$. When training on †, the student head weights corresponding to ‡ are also stationary; it remains for us to find equations of motion for $\mathbf{R}, \mathbf{Q}, \mathbf{U}$ and $\mathbf{h}^\dagger$, which we derive below. The equivalent derivations when training on teacher ‡ can be made by using the update in Eq. 5b instead.

C.2.1. ODE FOR $\mathbf{R}$

Consider the inner product of Eq. 5a (in the case of * = †) with $\mathbf{w}_n^\dagger$:

$$\mathbf{w}_k^{\mu+1}\mathbf{w}_n^\dagger - \mathbf{w}_k^\mu \mathbf{w}_n^\dagger = -\frac{\alpha_\mathbf{W}}{\sqrt{D}} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\mathbf{x}^\mu \mathbf{w}_n^\dagger \tag{C.9}$$

$$= -\alpha_\mathbf{W} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\rho_n^\mu \tag{C.10}$$

$$r_{kn}^{\mu+1} - r_{kn}^\mu = -\frac{\alpha_\mathbf{W}}{D} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\rho_n^\mu \tag{C.11}$$

If we let $\tau \equiv \mu/D$ and take the thermodynamic limit of $D \to \infty$, the time parameter becomes continuous and we can write:

$$\frac{dr_{in}}{d\tau} = -\alpha_\mathbf{W} h_i^\dagger \langle g'(\lambda_i)\Delta^\dagger \rho_n\rangle, \tag{C.12}$$

where we have re-indexed $k \to i$.

C.2.2. ODE FOR $\mathbf{Q}$

Consider squaring Eq. 5a (here we can simply use * to denote training on either teacher).

$$\mathbf{w}_k^{\mu+1}\mathbf{w}_i^{\mu+1} - \mathbf{w}_k^\mu \mathbf{w}_i^\mu = -\frac{\alpha_\mathbf{W}}{\sqrt{D}} h_i^{*\mu} g'(\lambda_i^\mu)\Delta^{*\mu}\mathbf{x}^\mu \mathbf{w}_k^\mu - \frac{\alpha_\mathbf{W}}{\sqrt{D}} h_k^{*\mu} g'(\lambda_k^\mu)\Delta^{*\mu}\mathbf{x}^\mu \mathbf{w}_i^\mu$$
$$+ \frac{\alpha_\mathbf{W}^2}{D} h_i^{*\mu} g'(\lambda_i^\mu) h_k^{*\mu} g'(\lambda_k^\mu)(\Delta^{*\mu}\mathbf{x}^\mu)^2 \tag{C.13}$$

$$= -\alpha_\mathbf{W} h_i^{*\mu} g'(\lambda_i^\mu)\Delta^{*\mu}\lambda_k^\mu - \alpha_\mathbf{W} h_k^{*\mu} g'(\lambda_k^\mu)\Delta^{*\mu}\lambda_i^\mu$$
$$+ \frac{\alpha_\mathbf{W}^2}{D} h_i^{*\mu} g'(\lambda_i^\mu) h_k^{*\mu} g'(\lambda_k^\mu)(\Delta^{*\mu}\mathbf{x}^\mu)^2 \tag{C.14}$$

$$q_{ki}^{\mu+1} - q_{ki}^\mu = -\frac{\alpha_\mathbf{W}}{D} h_i^{*\mu} g'(\lambda_i^\mu)\Delta^{*\mu}\lambda_k^\mu - \frac{\alpha_\mathbf{W}}{D} h_k^{*\mu} g'(\lambda_k^\mu)\Delta^{*\mu}\lambda_i^\mu$$
$$+ \frac{\alpha_\mathbf{W}^2}{D^2} h_i^{*\mu} g'(\lambda_i^\mu) h_k^{*\mu} g'(\lambda_k^\mu)(\Delta^{*\mu}\mathbf{x}^\mu)^2. \tag{C.15}$$

Performing the same reparameterisation of μ and the same thermodynamic limit, we get:

$$\frac{dq_{ik}}{d\tau} = -\alpha_\mathbf{W} h_i^* \langle g'(\lambda_i)\Delta^* \lambda_k\rangle - \alpha_\mathbf{W} h_k^* \langle g'(\lambda_k)\Delta^* \lambda_i\rangle + \alpha_\mathbf{W}^2 h_i^* h_k^* \langle g'(\lambda_i) g'(\lambda_k)\Delta^{*2}\rangle. \tag{C.16}$$

Note: in the limit, $(\mathbf{x}^\mu)^2 \to D$ since individual samples are taken from a unit normal. Hence the $1/D$ limit remains the same decay rate for each term.

C.2.3. ODE FOR $\mathbf{U}$

Consider the inner product of Eq. 5a (in the case of * = †) with $\mathbf{w}_p^\ddagger$:

$$\mathbf{w}_k^{\mu+1}\mathbf{w}_p^\ddagger - \mathbf{w}_k^\mu \mathbf{w}_p^\ddagger = -\frac{\alpha_\mathbf{W}}{\sqrt{D}} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\mathbf{x}^\mu \mathbf{w}_p^\ddagger \tag{C.17}$$

$$= -\alpha_\mathbf{W} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\eta_p^\mu \tag{C.18}$$

$$u_{kp}^{\mu+1} - u_{kp}^\mu = -\frac{\alpha_\mathbf{W}}{D} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\eta_p^\mu. \tag{C.19}$$

If we let $\tau \equiv \mu/D$ and take the thermodynamic limit of $D \to \infty$:

$$\frac{du_{ip}}{d\tau} = -\alpha_{\mathbf{W}} h_i^* \langle g'(\lambda_i)\Delta^* \eta_p \rangle. \tag{C.20}$$

C.2.4. ODE FOR $\mathbf{h}^*$

Here, we simply take the thermodynamic limit of Eq. 5b (for * = †):

$$\frac{dh_i^\dagger}{d\tau} = -\alpha_h \langle \Delta^\dagger g(\lambda_i) \rangle \tag{C.21}$$

D. Explicit Formulation

We can go one step further and write the right hand sides of the ODEs in terms of more concise integrals. Recall that for no noise

$$\Delta^{\dagger\mu} \equiv \sum_k h_k^{\dagger\mu} g(\lambda_k^\mu) - \sum_m v_m^\dagger g(\rho_m^\mu). \tag{D.1}$$

Substituting this term into the ODEs above gives us the expanded versions below:

$$\frac{dr_{in}}{d\tau} = -\alpha_{\mathbf{W}} h_i^\dagger \left\langle g'(\lambda_i) \left[\sum_k h_k^\dagger g(\lambda_k) - \sum_m v_m^\dagger g(\rho_m) \right] \rho_n \right\rangle; \tag{D.2}$$

$$\begin{aligned} \frac{dq_{ik}}{d\tau} = &-\alpha_{\mathbf{W}} h_i^\dagger \left\langle g'(\lambda_i) \left[\sum_j h_j^\dagger g(\lambda_j) - \sum_m v_m^\dagger g(\rho_m) \right] \lambda_k \right\rangle \\ &- \alpha_{\mathbf{W}} h_k^\dagger \left\langle g'(\lambda_k) \left[\sum_j h_j^\dagger g(\lambda_j) - \sum_m v_m^\dagger g(\rho_m) \right] \lambda_i \right\rangle \\ &+ \alpha_{\mathbf{W}}^2 h_i^\dagger h_k^\dagger \left\langle g'(\lambda_i) g'(\lambda_k) \left[\sum_j h_j^\dagger g(\lambda_j) - \sum_m v_m^\dagger g(\rho_m) \right]^2 \right\rangle; \end{aligned} \tag{D.3}$$

$$\frac{du_{ip}}{d\tau} = -\alpha_{\mathbf{W}} h_i^\dagger \left\langle g'(\lambda_i) \left[\sum_k h_k^\dagger g(\lambda_k) - \sum_m v_m^\dagger g(\rho_m) \right] \eta_p \right\rangle; \tag{D.4}$$

$$\frac{dh_i^\dagger}{d\tau} = -\alpha_{\mathbf{h}} \left\langle \left[\sum_k h_k^\dagger g(\lambda_k) - \sum_m v_m^\dagger g(\rho_m) \right] g(\lambda_i) \right\rangle. \tag{D.5}$$

Similarly to the I_2 integral defined in Eq. C.3, we further define:

$$I_3(d, f, h) = \langle g'(\zeta)\beta g(\gamma) \rangle, \tag{D.6}$$

$$I_4(d, e, f, h) = \langle g'(\zeta) g'(\iota) g(\beta) g(\gamma) \rangle; \tag{D.7}$$

where ζ, ι are local fields of the student with indices d, e; and β, γ can be local fields of either student or teacher with indices f, h. Substituting these definitions into the expanded ODE formulations gives:

$$\frac{dr_{in}}{d\tau} = \alpha_{\mathbf{W}} h_i^\dagger \left[\sum_m^M v_m^* I_3(i,n,m) - \sum_k^K h_k^\dagger I_3(i,n,k) \right]; \tag{D.8}$$

$$\begin{aligned}\frac{dq_{ik}}{d\tau} = {} & \alpha_{\mathbf{W}} h_i^\dagger \left[\sum_m^M v_m^\dagger I_3(i,k,m) - \sum_j^K h_j^\dagger I_3(i,k,j) \right] \\ & + \alpha_{\mathbf{W}} h_k^\dagger \left[\sum_m^M v_m^\dagger I_3(k,i,m) - \sum_j^K h_j^\dagger I_3(k,i,j) \right] \\ & + \alpha_{\mathbf{W}}^2 h_i^\dagger h_k^\dagger \left[\sum_{j,l}^K h_j^\dagger h_l^\dagger I_4(i,k,j,l) + \sum_{m,n}^M v_m^\dagger v_n^\dagger I_4(i,k,m,n) \right. \\ & \left. -2 \sum_j^K \sum_m^M v_m^\dagger h_j^\dagger I_4(i,k,j,m) \right];\end{aligned} \tag{D.9}$$

$$\frac{du_{ip}}{d\tau} = \alpha_{\mathbf{W}} h_i^\dagger \left[\sum_m^M v_m^\dagger I_3(i,p,m) - \sum_k^K h_k^\dagger I_3(i,p,k) \right]; \tag{D.10}$$

$$\frac{dh_i^\dagger}{d\tau} = \alpha_{\mathbf{h}} \left[\sum_m^M v_m^\dagger I_2(m,i) - \sum_k^K h_k^\dagger I_2(k,i) \right]. \tag{D.11}$$

This completes the picture for the dynamics of the generalisation error. It can be expressed purely in terms of the head weights and the I integrals. For the case of the scaled error function we can evaluate the I_2, I_3, and I_4 analytically meaning we have an exact formulation of the generalisation error dynamics of the student with respect to both teachers in the thermodynamic limit. Further details on the integrals can be found in App. E. The next chapter introduces the experimental framework that compliments the theoretical formalism presented above.

E. Gaussian Integrals under Scaled Error Function

In the derivations of App. C, we introduce a set of integrals over multivariate Gaussian distributions, labelled I_2, I_3 and I_4. They are defined as:

$$I_2(f,h) \equiv \langle g(\beta) g(\gamma) \rangle, \tag{E.1}$$

$$I_3(d,f,h) \equiv \langle g'(\zeta) \beta g(\gamma) \rangle, \tag{E.2}$$

$$I_4(d,e,f,h) \equiv \langle g'(\zeta) g'(\iota) g(\beta) g(\gamma) \rangle; \tag{E.3}$$

where ζ, ι are local fields of the student with indices d, e; and β, γ can be local fields of either student or teacher with indices f, h; and g is the activation function.

These integrals do not have closed form solutions for the ReLU activation. For the scaled error function however, they can all be solved analytically. They are given by:

$$I_2 = \frac{1}{\pi} \arcsin \frac{c_{12}}{\sqrt{(1+c_{11})(1+c_{22})}}; \tag{E.4}$$

$$I_3 = \frac{2c_{23}(1+c_{11}) - 2c_{12}c_{13}}{\sqrt{\Lambda_3}(1+c_{11})}; \tag{E.5}$$

$$I_4 = \frac{4}{\pi^2 \sqrt{\Lambda_4}} \arcsin \frac{\Lambda_0}{\sqrt{\Lambda_1 \Lambda_2}}; \tag{E.6}$$

where

$$\Lambda_0 = \Lambda_4 c_{34} - c_{23}c_{24}(1+c_{11}) - c_{13}c_{14}(1+c_{22}) + c_{12}c_{13}c_{24} + c_{12}c_{14}c_{23}; \tag{E.7}$$
$$\Lambda_1 = \Lambda_4(1+c_{33}) - c_{23}^2(1+c_{11}) - c_{13}^2(1+c_{22}) + 2c_{12}c_{13}c_{23}; \tag{E.8}$$
$$\Lambda_2 = \Lambda_4(1+c_{44}) - c_{24}^2(1+c_{11}) - c_{14}^2(1+c_{22}) + 2c_{12}c_{14}c_{24}; \tag{E.9}$$
$$\Lambda_3 = (1+c_{11})(1+c_{33}) - c_{13}^2; \tag{E.10}$$
$$\tag{E.11}$$

and where c is the relevant projected down covariance matrix.

F. Overlap Generation

In subsubsection 2.1.2, we investigate the effect of task similarity on forgetting. In our framework, the teachers act as tasks. From App. C, we know that the learning dynamics in the student can be fully described by the overlap parameters, which includes the teacher-teacher overlap matrix, V. For our investigation we need a method to generate teachers with specific overlaps; specifically— in the normalised teachers Ansatz, and for teachers with a single hidden unit—we perform simulations over the full range of V from 0 to 1. In this configuration we simply need a procedure to generate two N-dimensional vectors, $\mathbf{v}_1$, $\mathbf{v}_2$, with an angle θ between them such that:

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = \theta. \tag{F.1}$$

Fortunately there is a standard algorithm for this. First we define two vectors

$$\tilde{\mathbf{v}}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}; \quad \tilde{\mathbf{v}}_2 = \begin{pmatrix} \sin\theta \\ \cos\theta \end{pmatrix}.$$

Second, we generate an $N \times N$ orthogonal matrix, R. There is a standard sicpy implementation for this based on QR decomposition of a random Gaussian matrix[2].

Finally, multiply the first two columns of R with either vector to generate the rotated vectors:

$$\mathbf{v}_1 = R[:, 1:2] \cdot \tilde{\mathbf{v}}_1; \tag{F.2}$$
$$\mathbf{v}_2 = R[:, 1:2] \cdot \tilde{\mathbf{v}}_2. \tag{F.3}$$

G. Experiment Details

In this section we provide details of experimental procedures used to obtain the graphs and figures presented in this work.

In the ODE limit investigation, the following parameters were used:

- Input dimension = 10,000;
- Test set size = 50,000;
- SGD optimiser;
- Mean squared error loss;
- Teacher weight initialisation: normal distribution with variance 1;
- Student weight initialisation: normal distribution with variance 0.001;
- Student hidden dimension: 2;
- Teacher hidden dimension: 1;
- Learning rate: 1

In the mean-field limit investigation the following parameters were used:

- Input dimension = 15;

[2] SciPy Stats Module Docs

- Test set size = 25,000;
- SGD optimiser;
- Mean squared error loss;
- Teacher weight initialisation: normal distribution with variance 1;
- Student weight initialisation: normal distribution with variance 0.001;
- Student hidden dimension: 1000;
- Teacher hidden dimension: 250;
- Learning rate: 5

H. Forgetting vs. V at Multiple Intervals

In Fig. 3, we show the cross section of forgetting vs. V at a set of intervals after the task boundary. In Fig. 7, we show this cross-section at a greater range of time delays after the switch.

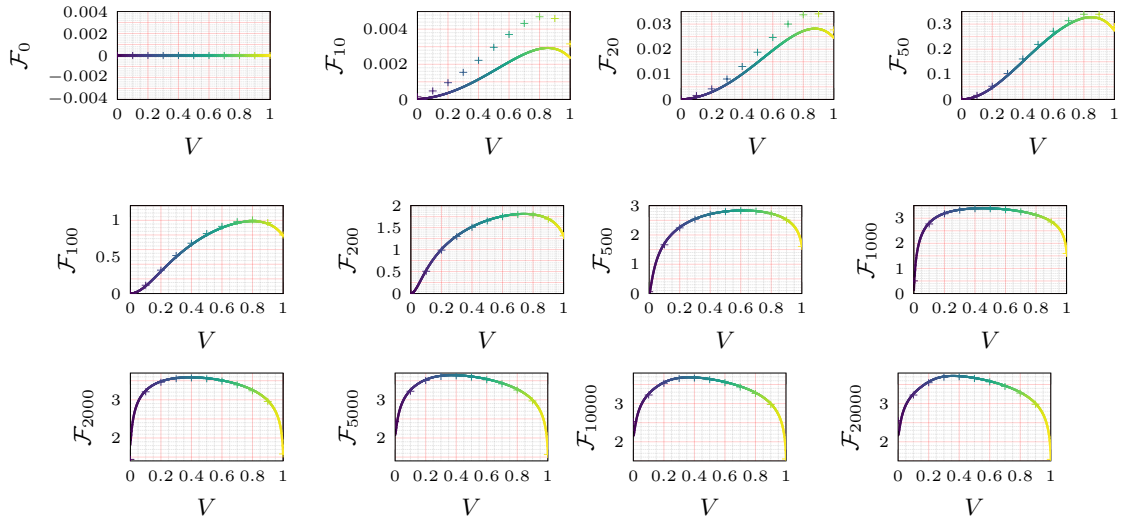

Figure 7. Aggregate forgetting, $\mathcal{F}_t$, vs. teacher-teacher overlap, V, at different time intervals post task-switch. A teacher-teacher overlap of 0 corresponds to orthogonal teacher weight vectors, whereas a teacher-teacher overlap of 1 corresponds to aligned teacher weight vectors. Forgetting is strongest for teachers that are intermediately correlated, while the student is relatively robust to forgetting for aligned or orthogonal teachers. The distribution of error changes moves significantly as time spent training on the new task increases.

I. Forgetting vs. Feature Similarity, ReLU Networks

This sections contains the same experiments as those presented in subsubsection 2.1.2, but for networks with ReLU nonlinearities. Fig. 8 shows for various values of V the generalisation error of the first teacher over time. Fig. 9 shows the cross sections of forgetting vs. V at various time intervals after the task switch.

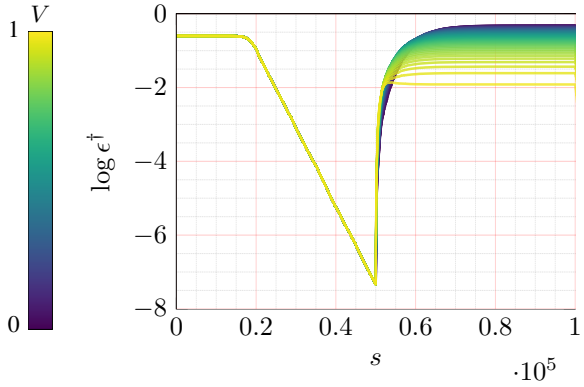

Figure 8. Generalisation error with respect to first teacher, $\log \epsilon^\dagger$, vs. timestep, s, for a range of teacher-teacher overlaps for ReLU networks. Task switches occur at steps 50,000 and 100,000. This plot is the ReLU equivalent of Fig. 3 in subsubsection 2.1.2

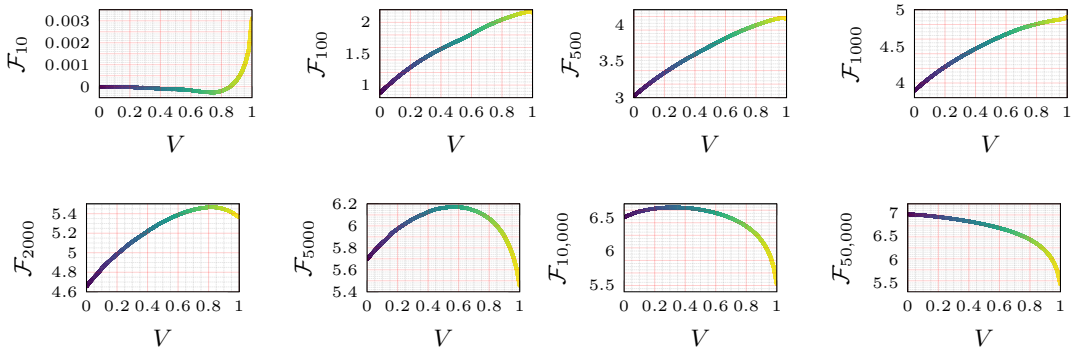

Figure 9. Aggregate forgetting, $\mathcal{F}_t$, vs. teacher-teacher overlap, V, at different time intervals post task-switch for ReLU networks. The distribution of error changes moves faster compared to the sigmoid case in subsubsection 2.1.2. By the second task switch, the function is monotonic.

J. Effect of Activation & Distribution Evolution

The learning dynamics and corresponding forgetting/transfer distributions for varying teacher-teacher overlaps presented above are for sigmoidal activation functions. In our investigations we found that different activation functions can have a strong impact on how the forgetting vs. teacher-teacher overlap distributions change over time. In particular, in the ReLU case, the distribution moves relatively quickly from a hump curve (seen in the sigmoidal case) to a monotonic function, where the higher overlaps lead to less forgetting. Detailed plots for the ReLU case are shown in App. I. Forgetting and transfer are not stationary attributes, hence the inclusion of a time component in our definitions of these quantities. The unsurprising observation that the distribution of forgetting over different overlaps changes as time progresses beyond the switch point is not discussed in previous research. The nature of this evolution and its contributing factors are worthy of further investigation.

K. First Task Convergence

The setting we work in throughout our experimentation is one in which good convergence has been achieved on the first task before the switch. Some of the observations we make are therefore conditional on this convergence. We show below in Fig. 10 one example of a phenomenon (higher rate of forgetting for greater task similarity) we observed in the main results that does not hold in settings where lesser convergence is achieved on the first task.

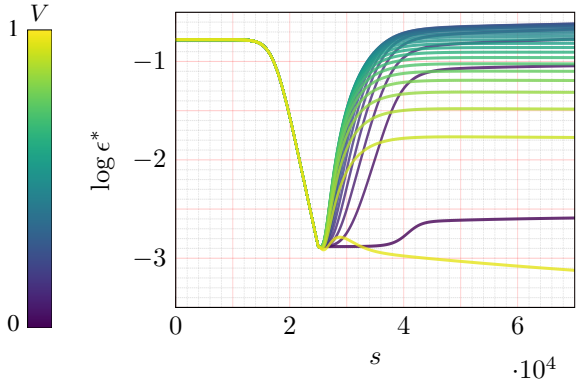

Figure 10. Generalisation error with respect to first teacher, $\log \epsilon^*$, vs. timestep, s, for a range of teacher-teacher overlaps, V. Here the task switch occurs relatively early—before convergence on the first task. Unlike in settings where better convergence has been achieved the initial rate of forgetting is not largest for highest overlap. In fact here there is a period of co-learning just after the task-switch.

L. Forgetting/Transfer Metrics Procedure (Mean-Field Limit)

In Fig. 5, we present metrics of forgetting and transfer for various task similarity configurations averaged over 50 random seeds. Specifically we give the initial rates, maxima, and long-time values. Here we provide details on how these are evaluated.

L.1. Initial Rate

Let $\tilde{s}$ be the training step at which the teacher switches. We approximate the initial rate of forgetting as:

$$\frac{1}{N}\sum_{i=1}^{N} \epsilon^{\dagger}|_{\tilde{s}+i} - \epsilon^{\dagger}|_{t=\tilde{s}+i-1}, \tag{L.1}$$

where N is the number of steps over which we take the average change ($N = 20$ for experiments shown in Fig. 5). Since we are not using the ODE solutions but pure simulation of the mean-field limit in Fig. 5, such a sampling is necessary to accurately approximate the rates. Likewise the initial rate of transfer is computed via:

$$\frac{1}{N}\sum_{i=1}^{N} \epsilon^{\ddagger}|_{\tilde{s}+i-1} - \epsilon^{\ddagger}|_{t=\tilde{s}+i}. \tag{L.2}$$

L.2. Maxima

The maximum forgetting and transfer amounts are computed with

$$\max_{t}(\epsilon^{\dagger}|_{\tilde{s}+t}) - \epsilon^{\dagger}|_{\tilde{s}} \quad \text{and} \quad \epsilon^{\dagger}|_{\tilde{s}} - \min_{t}(\epsilon^{\ddagger}|_{\tilde{s}+t}). \tag{L.3}$$

L.3. Long-Time Limit

Initially we computed the long-time limits simply as the differences in generalisation error at the end of training with those at the switch point. However, for forgetting we needed to adjust this procedure slightly. Fig. 11 shows a run associated with a single task configuration in the mean-field limit—in particular, this run is for tasks with full feature overlap.

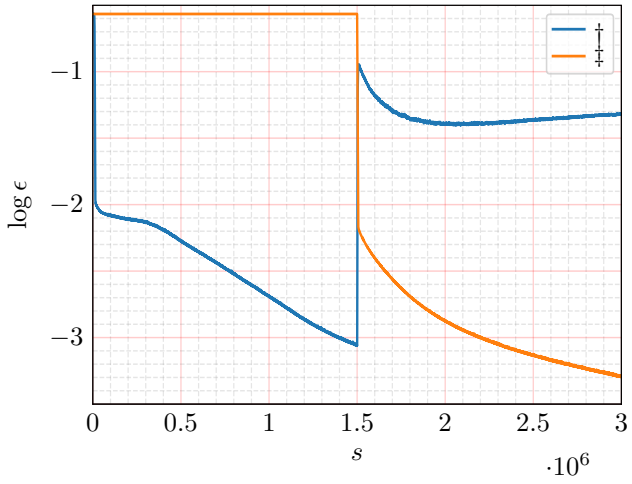

Figure 11. Generalisation errors, $\log \epsilon$, vs. training step, s for both teacher 1 (†) and teacher 2 (‡) for the mean-field limit with full feature similarity between teachers. In the second task phase, there is sharp initial forgetting. This is followed by a period of co-learning. Then at around two million steps there is a second turn of forgetting. This corresponds with the point at which the performance on the second task matches the best performance attained by the student with respect to the first teacher in the first phase.

M. $\tilde{V} = 0$ Row in Mean-Field Limit

We noted in Fig. 5 that the orthogonal readout row, $\tilde{V}$, displays similar trends to the results of varying the feature similarity in the ODE limit. Here we show more details plots from the row beyond the coarse heatmap in Fig. 5. Fig. 12 shows cross sections of forgetting vs. α at different intervals after the switch. They are the equivalent plots of Fig. 3 but for the orthogonal readout row runs of Fig. 5. They show that as for the feature similarity variation in the ODE limit, there is a non-mnotonic relationship between similarity and forgetting such that the intermediate similarity is worst. The development of the shape of the cross-section is also similar. Trivially it begins flat. The non-monoticity is sharpest at intermediate intervals after the switch, and in the long-time limit flattens out again with a wide peak and very little forgetting for large overlap.

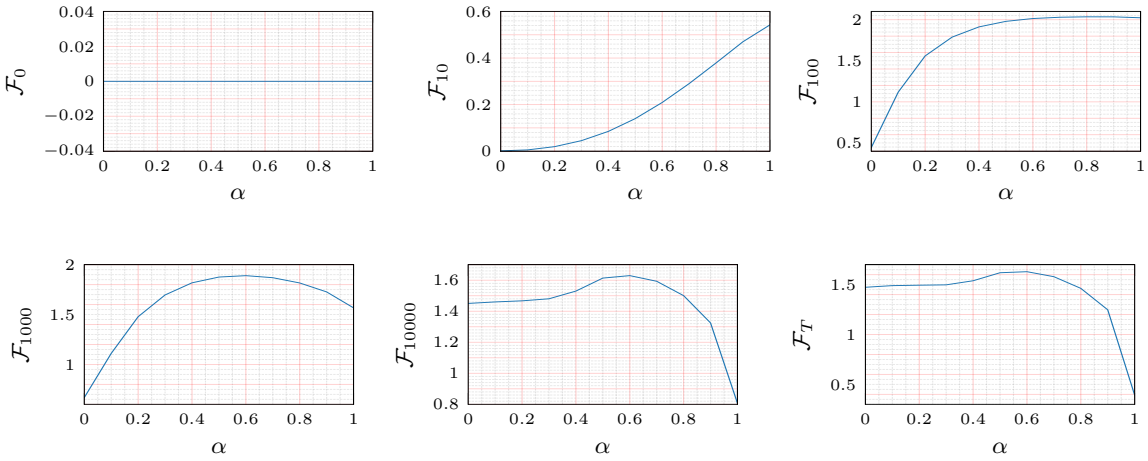

Figure 12. Aggregate forgetting, $\mathcal{F}_t$, vs. teacher-teacher feature overlap, α, for constant zero readout overlap, $\tilde{V} = 0$ in the mean-field limit, at different time intervals post task-switch.

N. Readout Bias on Feature Solution

One of the interesting results we found from the experiment shown in Fig. 5 was that for full feature overlap there was still variation in transfer ability for different readout similarities. After the switch in our multi-head student setup, the student is given a new set of randomly initialised head weights. The previously learned readout weights for the first task are (as far as the transfer ability is concerned) discarded. This newly initialised student head will be (approximately) orthogonal to all of the second teacher head weights, regardless of the relationship between the second teacher head weights and the first teacher head weights. Despite this, there is better transfer for the tasks where there is overlap in the teacher readouts. We hypothesis that this is due to a bias in SGD dynamics: during the first task phase, the local minimum that the solution finds within the feature space is biased by the readout weights it is concurrently trying to optimise. Taking an extreme example, suppose you have two hidden nodes and teacher 1 has readout weight = 1 on node 1, and 0 on node 2. While training on task 1, the network will not learn the input-to-hidden node 2 weights since this node does not impact the output. Therefore there will be a transfer cost if the second task relies on both nodes, which arises from the requirement to learn the input-to-hidden weights that were unimportant for task 1. We verify this idea empirically by tracking the movement of the feature weights after the task switch for different readout similarities. The results are shown in Fig. 13 and demonstrate that the feature weights move more (further away from the solution found for task 1, which has identical features to task 2) for task configurations with lower readout similarity.

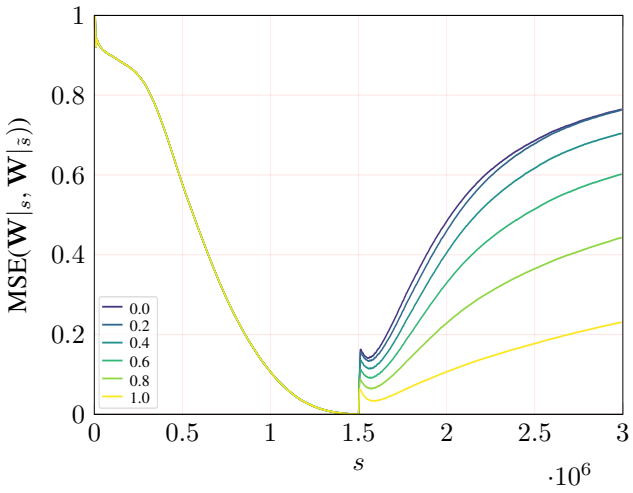

Figure 13. (Normalised) mean squared error between the student feature weights at a given step of training and the student feature weights at the switch point, $\mathbf{W}|_s, \mathbf{W}|_{\tilde{s}}$), vs. training step, s for full feature similarity and various readout similarity configurations in the mean-field limit. Trivially the MSE is 0 at the switch. After the switch, despite moving onto a new teacher with the same features as the first teacher, the student feature weights move. However they move more for task configurations in which the second readout weights are more dissimilar from the first.