Maslow's Hammer for Catastrophic Forgetting: Node Re-Use vs Node Activation

Sebastian Lee [1 2 3] **Stefano Sarao Mannelli** [2 3] **Claudia Clopath** [1 2] **Sebastian Goldt** [4] **Andrew Saxe** [2 3 5]

Abstract

Continual learning—learning new tasks in sequence while maintaining performance on old tasks—remains particularly challenging for artificial neural networks. Surprisingly, the amount of forgetting does not increase with the dissimilarity between the learned tasks, but appears to be worst in an intermediate similarity regime.

In this paper we theoretically analyse both a synthetic teacher-student framework and a real data setup to provide an explanation of this phenomenon that we name Maslow's hammer hypothesis. Our analysis reveals the presence of a trade-off between node activation and node re-use that results in worst forgetting in the intermediate regime. Using this understanding we reinterpret popular algorithmic interventions for catastrophic interference in terms of this trade-off, and identify the regimes in which they are most effective.

1. Introduction

Artificial neural networks have reached astonishing performance in a number of different applications (Silver et al., 2016; Rajkomar et al., 2019; Devlin et al., 2019; Tunyasuvunakool et al., 2021), but they tend to perform poorly when they have to solve a sequence of learned tasks (Kemker et al., 2018). The ineffectiveness of deep learning algorithms in this learning paradigm—known as continual learning or lifelong learning—is strikingly different from observations in human and animal learning, where tasks can effectively be learned sequentially and interference is a rarity (Barnett & Ceci, 2002; Calvert et al., 2004; Mareschal et al., 2007; Pallier et al., 2003). Neuroscientists and psychologists have been interested in the mechanisms underpinning this ability for some time (McCloskey & Cohen, 1989; Cichon & Gan, 2015; Yang et al., 2014; Flesch et al., 2018); more recently the engagement of the machine learning community with this paradigm has also grown as the focus shifts from performance on single tasks to distributions of tasks and learning from the real world (Parisi et al., 2019).

The key difficulty of continual learning is avoiding so-called *catastrophic forgetting* or *catastrophic interference* (McCloskey & Cohen, 1989; Ratcliff, 1990), the phenomenon of deteriorating performance on earlier tasks when learning later tasks. Humans are very good continual learners, and various biological mechanisms have been proposed to account for the brain's ability to combat forgetting (McClelland et al., 1995). On the other hand artificial systems, in particular neural networks trained with gradient descent algorithms, suffer badly from catastrophic forgetting (Goodfellow et al., 2014). This has prompted research into methods to augment vanilla gradient descent with additional elements specifically aimed at reducing forgetting (Parisi et al., 2019), including some that take inspiration form the aforementioned biological mechanisms (Hinton & Plaut, 1987; Robins, 1995; Gepperth & Karaoguz, 2016; Rebuffi et al., 2017).

In addition to this algorithmic line of research, there is growing interest in understanding *why* forgetting affects deep learning so severely and what the main drivers of forgetting are. Ramasesh et al. (2021) performed a series of systematic experiments in a range of architectures and training setups and made the counterintuitive observation that catastrophic forgetting is worst between tasks of *intermediate* similarity. Lee et al. (2021) then analysed the impact of task similarity on continual learning in a solvable model of two-layer neural networks and found the same non-monotonic relationship between task similarity and forgetting. Despite these results, the precise mechanism that makes intermediate task similarity the worst has remained unclear.

Here, we describe a possible mechanism that drives catastrophic forgetting in two-layer neural networks trained in a teacher-student setup. Our main contributions can be summarised as follows:

[1]Imperial College, London, UK [2]Sainsbury Wellcome Centre, UCL [3]Gatsby Computational Neuroscience Unit, UCL [4]International School of Advanced Studies (SISSA), Trieste, Italy [5]CIFAR Azrieli Global Scholars program, CIFAR, Toronto, Canada. Correspondence to: Sebastian Lee <sebastian.lee14@imperial.ac.uk>, Andrew Saxe <a.saxe@ucl.ac.uk>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

- Maslow's hammer hypothesis to explain the observation that intermediate similarity is worst for forgetting (Ramasesh et al., 2021; Lee et al., 2021) in terms of a trade-off between node (hidden unit) re-use and node activation;
- Evidence from both the student-teacher framework and a data-mixing image classification paradigm to support the Maslow's hammer hypothesis;
- An empirical study of how various methods of alleviating forgetting impact the relationship between task similarity and forgetting;
- Observation of '*catastrophic slowing*', whereby despite tremendous advantages in aligned and orthogonal task settings, interleaving can be inferior to regularisation methods in intermediate similarity regimes.

Further related work

Recent investigations into theoretical questions related to continual learning include work by Mirzadeh et al. (2021), who study the relationship between network width, depth and forgetting; showing that wider and shallower networks are less affected by forgetting. Bell & Lawrence (2021) take inspiration from methodologies in psychology to design tasks aimed at investigating catastrophic forgetting, including in relation to loss surfaces and the interplay between semantic and perceptual information. Meanwhile Shen et al. (2021) gained interesting insights into algorithmic components of two-layer neural circuits—not unlike those studied here—that allow fruit flies to mitigate interference, including sparse coding and associative learning. Arguably closest in nature to our work are those of Asanuma et al. (2021) and Doan et al. (2021), who also investigate the effect of task similarity on forgetting but in linear regression and Neural Tangent Kernel (NTK) (Jacot et al., 2018) regimes respectively.

A related theoretical line of research concerns transfer learning where the focus is not on forgetting, but on the boost that features learned on upstream tasks can provide new tasks (Tan et al., 2018). Dhifallah & Lu (2021) & Gerace et al. (2022) have analysed, on single-layer and two-layer networks respectively, the effect of similarity between tasks and data scarcity on performance in the downstream task using methods similar to ours.

On the more applied side of work on continual learning, concerned with developing algorithms to combat forgetting in neural networks, there is a larger body of literature (see e.g. Parisi et al. (2019) for a review). Methods can broadly be split into three categories: regularisation (Zenke et al., 2017; Li & Hoiem, 2017; Kirkpatrick et al., 2017); dynamic architectures (Rusu et al., 2016; Draelos et al., 2017; Zhou et al., 2012); and replay (McClelland et al., 1995; Shin et al., 2017). A more recent set of approaches concerns itself with explicitly learning modular representations for compositionality (Mendez & EATON, 2021; Veniat et al., 2021; Ostapenko et al., 2021); our investigation into node specialisation connects naturally to some of these concepts. In this work we look at aspects of each of these method families: insofar as our setups have separate heads for each task, we implicitly consider adaptive architectures; in Sec. 4 we explicitly investigate how Elastic Weight Consolidation (EWC) and interleaved replay affect the relationship between task similarity and forgetting.

2. Continual Learning Setup

In this work we consider two paradigms to study continual learning: a synthetic framework using the teacher-student model, and a real data framework where similarity is parameterized by a mixing parameter.

Teacher-Student Framework. The key idea of the teacher-student setup is to train a neural network, the student, on a data set generated by taking random inputs and propagating them through a fixed neural network with random weights called the teacher (Gardner & Derrida, 1989; Seung et al., 1992; Engel & Van den Broeck, 2001). We will consider two-layer networks with output $\phi(\mathbf{x}; \mathbf{W}, \mathbf{v}) = \sum_{l=1}^{L} \mathbf{v}_l g\left(\mathbf{w}_l \cdot \mathbf{x} / \sqrt{D}\right)$; where D is the input dimension, L is the number of hidden units, $\mathbf{W} \in \mathbb{R}^{L \times D}$ are the first layer weights, $\mathbf{v} \in \mathbb{R}^{L}$ are the second layer weights, g is the activation function, and $\mathbf{x} \in \mathbb{R}^{D}$ is the input vector. These inputs are sampled i.i.d. (independent and identically distributed) from the standard normal distribution.

While the framework allows for any number of tasks, for concreteness we consider training the student on a succession of two tasks (which we denote throughout by † and ‡). In the i^{th} phase of training (i.e. training on the i^{th} task), the supervision labels for the student are generated from the i^{th} teacher by $y^i = \phi(\mathbf{x}; \mathbf{W}^i, \mathbf{v}^i)$ and the student outputs are given by $\hat{y}^i = \phi(\mathbf{x}; \mathbf{W}, \mathbf{h}^i)$. Training is performed with Stochastic Gradient Descent (SGD) on the squared error $(y^i - \hat{y}^i)^2$. Note that the student shares first layer weights $\mathbf{W}$ across tasks but has separate head weights $\mathbf{h}^i$ for each task. A sketch of this setup is shown in Fig. 1a.

The key quantity that we would like to compute is the generalisation error of the student with respect to the i^{th} teacher,

$$\epsilon^i(\mathbf{W}, \mathbf{h}^i, \mathbf{W}^i, \mathbf{v}^i) \equiv \frac{1}{2} \left\langle \left[\phi(\mathbf{x}; \mathbf{W}^i, \mathbf{v}^i) - \phi(\mathbf{x}; \mathbf{W}, \mathbf{h}^i)\right]^2 \right\rangle . \quad (1)$$

Note that due to the separate heads, the generalisation errors are well-defined with respect to both teachers regardless of which is currently providing the labels. From these generalisation errors, one can further define quantities analogous to forgetting and transfer from one teacher/task to the next as differences in generalisation error.

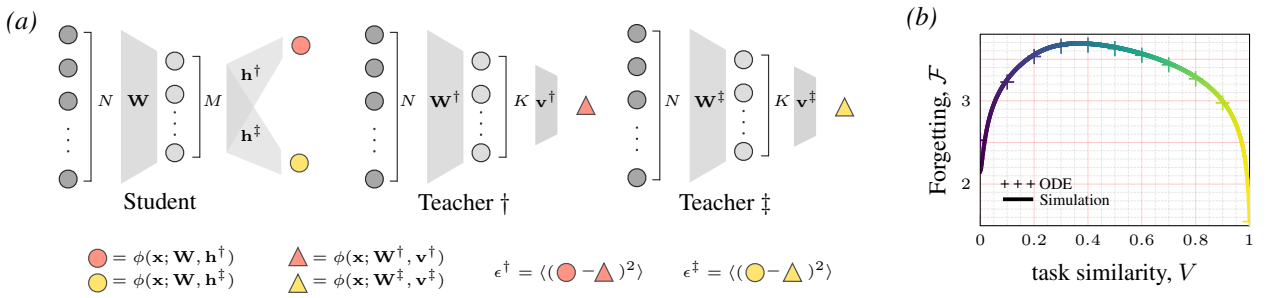


Figure 1. **Intermediate task similarity leads to worst catastrophic forgetting in the teacher-student setup** *(a)* Schematic of student teacher framework for continual learning. The student is over-parameterised with respect to teachers. Labels for training student are generated by teacher † in first phase of training and teacher ‡ in second phase. *(b)* Forgetting given by the difference between generalisation error on task 1 at the switch and (in this plot) after 10,000 steps as function of teacher-teacher similarity V (cf. Sec. 3.2). Crosses show solution from ODEs while solid line shows simulations.

The advantage of the teacher-student setup is that by providing full control over the input distribution, the similarity between tasks can be precisely tuned by controlling the relationship between teacher weights. Furthermore, we can give the student the right number of parameters to learn all teachers perfectly, at least in principle. Finally, the dynamics of the student in this setup can be solved exactly, yielding an ODE that describes its average dynamics (shown for single teacher-student by Saad & Solla (1995), and later for continual learning by Lee et al. (2021)). A key observation from this framework is that intermediate task similarity is worst for forgetting; we reproduce this in Fig. 1b (c.f. Fig. 3 in Lee et al. (2021) and App. A for details). In this work, we propose a mechanism responsible for this behaviour, and investigate the impact of various methods for combating forgetting on this relationship.

Data-Mixing Framework. To probe how well our findings translate to more realistic data distributions, we complement the teacher-student framework with a data-mixing approach similar to that introduced by Ramasesh et al. (2021). This procedure gives a notion of control over the similarity between any pair of tasks in that the input distribution is composed of a mixture between two separate datasets, where the mixing factor determines the similarity. More specifically, consider two separate datasets with equal cardinality in which inputs and outputs have the same dimensions across both datasets, i.e. $\mathcal{D}_1 = \{\mathbf{x}_i^1, y_i^1\}_{i=1}^N$ and $\tilde{\mathcal{D}}_2 = \{\tilde{\mathbf{x}}_i^2, \tilde{y}_i^2\}_{i=1}^N$. We can control the 'similarity' between two successive tasks by first training on $\mathcal{D}_1$, followed by a mixture between $\mathcal{D}_1$ and $\tilde{\mathcal{D}}_2$. For a given mixing factor α, this second dataset is given by

$$\mathcal{D}_2^\alpha = \{\alpha \mathbf{x}_i^1 + (1-\alpha)\tilde{\mathbf{x}}_i^2, \alpha y_i^1 + (1-\alpha)\tilde{y}_i^2\}_{i=1}^N. \quad (2)$$

Under this protocol, $\alpha = 0$ corresponds to a completely new dataset and an entirely new task, whereas $\alpha = 1$ corresponds to continuing to train on the first task.

3. Intermediate Task Similarity

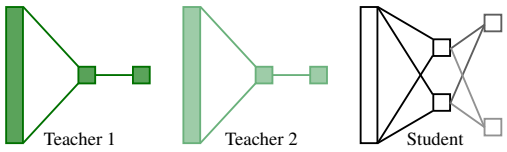
Although numerous independent studies have found non-monotonic relationships between task similarity and forgetting in artificial neural networks (Ramasesh et al., 2021; Lee et al., 2021; Asanuma et al., 2021), a convincing explanation for this result is still missing. Here, we propose such a mechanism, which we call Maslow's hammer hypothesis. The starting point is to think about how individual nodes in the hidden layers of two-layer networks are re-purposed during continual learning. We first outline the intuition behind the hypothesis before presenting supporting evidence from both the teacher-student setup and networks trained on image data.

3.1. Maslow's Hammer Hypothesis

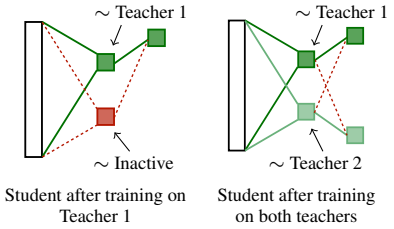
The trade-off between node re-use and node activation finds an analogy in a well-known cognitive bias in psychology: in the words of Abraham Maslow, "*[...] it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail*" (Maslow, 1966). In other words: if we have a nail we can (and should) use a hammer, but for a screw we really need a different tool and should avoid using the hammer.

This phenomenon illustrates the choice that the student network makes in learning the new task. Consider the simplest case of a student with two hidden units trained on a pair of teachers with a single hidden unit each. Since the student has one set of head weights for each teacher, it has the capacity to achieve zero test error on both teachers at the same time, cf. Fig. 2a.

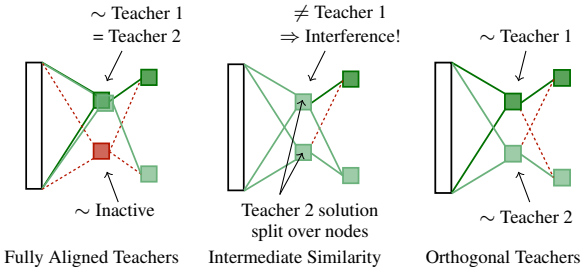
During the first task, we expect the student to learn the first teacher. In doing so, we *assume* a high degree of specialisation in the student whereby a subset of units in the network being trained become very important for the task while others remain inactive or unimportant. In this specific case, this results in the student using one node to learn the first teacher



(a) Student-Teacher Setup



(b) Minimal interference solution



(c) Characteristic solutions found by vanilla SGD

Figure 2. **Maslow's hammer hypothesis sketch** These diagrams outline the hypothesis for why intermediate task similarity is worst for forgetting using a minimal student-teacher example i.e. teachers with a single hidden unit and a student with two hidden units—as shown in *(a)*. *(b)* shows an ideal solution; here one student node specialises to the first teacher in the first phase of training and the second student node specialises to the other in the second phase without changing the first node. *(c)* shows the typical solutions found by the student when trained with vanilla SGD. For orthogonal teachers the student finds a solution in the same way as outlined for optimality in *(b)*. For fully aligned teachers, the student mostly re-uses the previously specialised node and ignores its spare capacity. In the intermediate region the student attempts a hybrid solution using both nodes, which leads to most interference.

and leaving the second node virtually inactive. We will show that this is the case in both the teacher-student setup and on the data mixture (cf. Fig. 3, Fig. 4 and App. B)

After the switch point there are three ways in which the student can learn the second teacher: *re-use* of the node that specialised to the first teacher, *activation* of the second (previously inactive) node, or a hybrid of the two, cf. Fig. 2c. In order to minimise forgetting, the student could use the inactive units to learn the second teacher and leave the node that has specialised on the first teacher untouched, as schematised in Fig. 2b. If the tasks are very related, however, it may be convenient to re-use previously activated nodes and leverage the features extracted in the first task, as in transfer learning (Tan et al., 2018). If the tasks are very dissimilar, the orthogonal teacher regime, the student network typically chooses to begin using its inactive node leaving the specialised one fairly intact, thus guarding against catastrophic interference. The intermediate case represents the most difficult case for the student: the previously specialised node will be somewhat aligned to the second teacher, so there would be some transfer benefit to fine tuning the orientation of this node (*re-use*), and making up any remaining difference with the previously inactive node. However under such a policy, unlike for the fully aligned teacher case, this would result in interference since the specialised node is moved. In sum, the Maslow's hammer hypothesis states that gradient descent dynamics bias towards re-using nodes when tasks are more similar (using a hammer for increasingly nail-like objects) and towards activation when tasks are more dissimilar (finding a different tool for decreasingly nail-like objects). This is most damaging when tasks are somewhat related, akin to breaking a screw when attempting to use a hammer.

3.2. Evidence from the Teacher-Student Framework

We first focus on the teacher-student setting where the model assures precise control. The framework detailed in Sec. 2 can be analysed exactly. Indeed, a long line of work has shown that, as the input dimension tends to infinity, the generalisation error concentrates and can be understood purely in terms of so-called 'order parameters' of the system (Mézard et al., 1987; Engel & Van den Broeck, 2001). Here, the concentration hypothesis will be taken as a working hypothesis and verified numerically by comparing theory and simulations.

We focus on the simplest setup to exhibit catastrophic forgetting, a two-task setting with a single task switch. Each teacher has a single hidden node and an output weight of norm one; the student has two hidden nodes. Among the crucial order parameters in the two-layer teacher-student scenario are the teacher-student overlaps $r_{km} \equiv \frac{1}{D}\mathbf{w}_k^T\mathbf{w}_m^\dagger$ and $u_{kp} \equiv \frac{1}{D}\mathbf{w}_k^T\mathbf{w}_p^\ddagger$, which measure the alignment between the weight of the m^{th} (p^{th}) teacher $\dagger$ ($\ddagger$) node and the k^{th} student node. At the beginning of training, the random teacher weights and the randomly initialised student vector have very little overlap; throughout training, the student will improve its alignment with the teacher providing the labels, and hence its test error. Another order parameter is the overlap between student weights, $q_{k\ell} \equiv \frac{1}{D}\mathbf{w}_k^T\mathbf{w}_\ell$, where the diagonal elements give the student node norms. The final crucial order parameter for our continual learning analysis, is the overlap between the first-layer weights of the two teachers, $v_{mp} = \frac{1}{D}(\mathbf{w}_m^\dagger)^T\mathbf{w}_p^\ddagger$, which we abbreviate to V (see App. C & App. E for details). This order parameter describes the task similarity: orthogonal tasks ($V = 0$) cor-

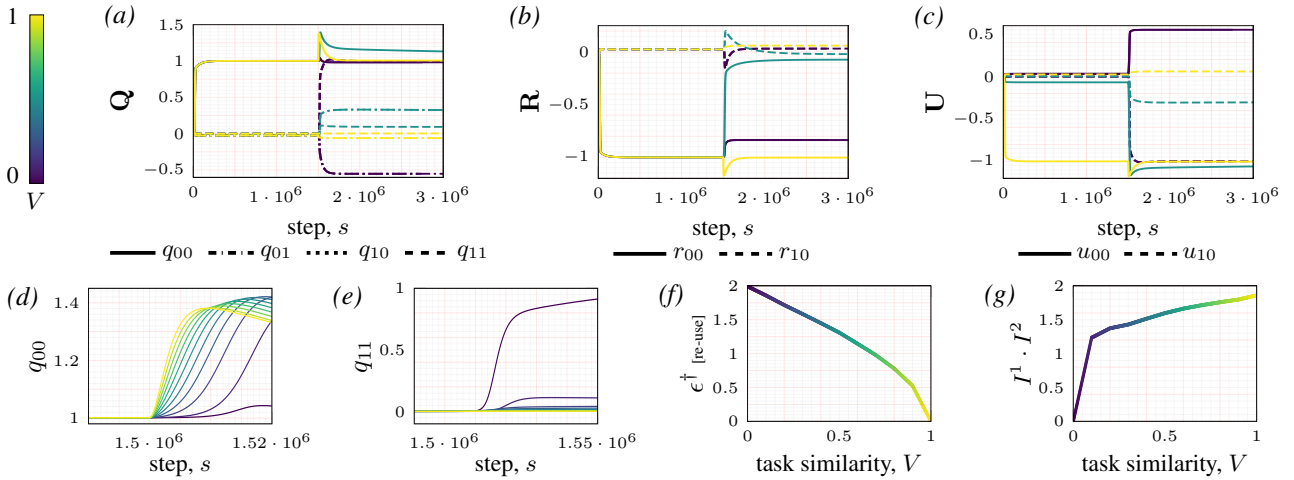


Figure 3. **Maslow's hammer in teacher-student setup** *(a)* Student self-overlap, $\mathbf{Q}$ *(b)* student-teacher 1 overlap, $\mathbf{R}$ and *(c)* student-teacher 2 overlap, $\mathbf{U}$ plotted vs. step s for a two layer student network on two successive teachers with single hidden unit (switch at $1.5e^6$). Yellow lines show trajectories for highly aligned teachers $V = 1$, purple lines for orthogonal teachers $V = 0$, while turquoise represents intermediate similarity. **Re-use vs. activation tendency** *(d)* Specialised student node norm q_{00} vs. time-step s around the switch point. Rate of movement from the fixed point *increases* monotonically as a function of teacher-teacher similarity, demonstrating the tendency for node re-use when tasks are highly similar. *(e)* Inactive student node norm q_{11} vs. time-step s around the switch point. Rate of movement from the fixed point *decreases* monotonically as a function of teacher-teacher similarity, demonstrating the tendency for node activation when tasks are highly dissimilar. *(f)* Proxy for asymptotic generalisation error with respect to first teacher under full re-use, $\epsilon^{\dagger}$ [re-use] (see Eq. 3) vs. teacher-teacher similarity V. The cost of node re-use is highest for orthogonal tasks. *(g)* dot product of first task importance vector with second task importance vector $I^1 \cdot I^2$ vs. teacher-teacher similarity V. Higher values correspond to similar nodes being important for the second and first task, thus indicating a bias towards node re-use. This plot is equivalent to Fig. 4c.

respond to independent teacher weights, whereas perfectly similar tasks ($V = 1$) have identical teacher weights up to permutations of the second-layer weights.

These order parameters obey a closed set of differential equations that describes their dynamics when training the student using SGD as described in Sec. 2. These were first derived by Riegler & Biehl (1995) and Saad & Solla (1995), and recently shown to be asymptotically exact by Goldt et al. (2019). These methods have been used to explore a broad range of phenomena in neural networks, see Saad (2009) for a summary of early work and Yoshida & Okada (2019); Goldt et al. (2020); Refinetti et al. (2021); Saglietti et al. (2021) for recent results. Here, we follow the derivation of similar equations for continual learning provided by Lee et al. (2021). In Fig. 3 we show the evolution of the solution to these ODEs for a student trained on two successive teachers with several values of the similarity parameter V (see App. C for details) These plots show evidence for the Maslow's hammer hypothesis outlined in Sec. 3.1. It is clear from Fig. 3a. and Fig. 3b. that there is strong specialisation in the student network. The magnitude of one student node is close to 1, and that node is almost fully aligned with the first teacher before the switch. Meanwhile the other node is essentially inactive before the switch. In order to minimise the amount of forgetting, the specialised node should not move after the switch (as per Fig. 2b). However after the switch there is movement in both nodes. Let us consider different levels of similarity separately:

Fully aligned case (yellow lines): there is an initial phase of movement in the specialised node before a reversion to the solution found for the first task. Meanwhile the previously inactive node remains largely inactive. Although this solution does not use the spare capacity available to the student (outside the initial transient), it is very close to optimal behaviour in terms of forgetting.

Fully orthogonal case (purple lines): there is only a minimal deviation of the specialised node before it reverts to the solution found for the first task. On the other hand, there is complete activation and alignment to the second teacher of the second node. This can be seen in the darkest dashed line moving close to 1 in Fig. 3a, and the darkest dashed line in Fig. 3c moving close to -1 (sign is flipped by learned head weight). This solution is very close to the optimal one proposed by Fig. 2.

Intermediate case (turquoise lines): between the aligned and orthogonal cases, the student does a combination of re-using the previously specialised node and activating the previously inactive node. The former can be seen in the movement away from 1 in the solid lines in Fig. 3a and the movement away from -1 in the solid lines in Fig. 3b. The latter can be seen in the movements away from 0 of the dashed lines in Fig. 3a and the dashed lines in Fig. 3c.

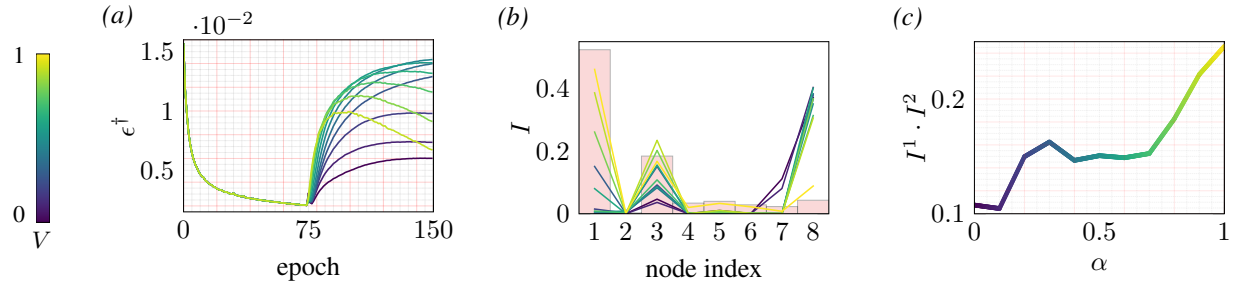


Figure 4. **Evidence for Maslow's Hammer in Data Mixing Setup** *(a)* Test error on first task $\epsilon^\dagger$ vs. epoch on two Fashion MNIST binary classification tasks where intermediate similarity between tasks is worst for forgetting; *(b)* Node importance I vs. node index. Bars show first task importance I^1 at switch point, while lines show second task importances I^2 at end of training; *(c)* dot product of first task importance vector with second task importance vector $I^1 \cdot I^2$ vs. mixing parameter α. Higher values correspond to similar nodes being important for the second and first task, thus indicating a bias towards node re-use. This plot is equivalent to Fig. 3g.

Re-use vs. activation tendency. An important aspect of this explanation is that more similar tasks will bias the network towards re-use whereas more orthogonal tasks will bias the network towards new activation. This is not obvious *a priori* and warrants closer inspection. In Fig. 3d we show the trajectory of the specialised student node norm around the switch point. Immediately after the switch there is a clear monotonic relationship between the teacher-teacher similarity and rate of movement in q_{11}. On the other hand, Fig. 3e shows the inverse relationship for movement away from 0 in the norm of the inactive student node. To complete this part of the picture, Fig. 3f shows the asymptotic generalisation error of the student with respect to the first teacher under a complete re-use scheme, which we compute retrospectively via:

$$\epsilon^\dagger_{\text{[re-use]}} = \frac{1}{2}\left\langle [\phi(\mathbf{x}; \mathbf{W}^\dagger, \mathbf{v}^\dagger) - \phi(\mathbf{x}; \mathbf{W}^\ddagger, \mathbf{h}^{\dagger *})]^2 \right\rangle, \quad (3)$$

where $\mathbf{W}^\dagger$ is the first teacher feature weights, $\mathbf{v}^\dagger$ is the first teacher head, $\mathbf{W}^\ddagger$ is the second teacher feature weights and $\mathbf{h}^{\dagger *}$ is the component of the student's first head weight reading from the specialised node at the switch point. There is another clear decreasing monotonic relationship between cost in re-using the node and task similarity. Together these plots show the re-use and activation tendencies of the student in various similarity regimes, as well as the costs associated with these tendencies.

3.3. Evidence from Data Mixing Framework

In this section we use the protocol described in Eq. 2 to supplement the analysis in the synthetic framework with real data experiments. It is worth bearing in mind the following implications of changing settings: (i) While previously task similarity was defined over input-output mappings and the input distribution was constant, here the similarity parameter influences both input-output mappings and input distributions; (ii) Since the tasks themselves share many features (e.g. edges for image datasets), this contributes additional similarity such that even no mixing ($\alpha = 0$) will give similar tasks to some extent.

In the teacher-student framework we can understand clearly how the student solution relates to the task given by the teacher network from the overlap matrices. In the standard supervised learning setting there is no analogue. Instead we define an empirical measure of node '*importance*' that we use to investigate how different nodes in the network contribute to forgetting. For a given node i, we define its importance I_i^t in relation to some task t to be the change in test error when the output of that node is masked (see App. F for details). If a node is important to the network for a given task, the error will increase substantially when this node is masked and I_i^t will be high.

In Fig. 4b we see that for a two-layer network trained on a Fashion MNIST (Xiao et al., 2017) binary classification task, one or two nodes dominate in terms of importance at the switch point. Any forgetting that occurs after the switch point will hence be dominated by the behaviour of these nodes. Empirically, we then observe that these nodes remain more important for the second task when the task similarity is higher (see Fig. 4b for a single seed). To visualise this more generally, we consider the dot product of the vector of node importances for the first task at the switch point I^1 with the vector of node importances for the second task at the end of training on both tasks I^2. If similar nodes are important for both tasks (re-use), this quantity is high; while if different nodes are important (activation), it is low. Fig. 4c shows that, just as in the teacher-student (see Fig. 3g), $I^1 \cdot I^2$ generally increases as a function of similarity in the data-mixing setup (see App. G for details on statistics).

4. Methods for Combating Forgetting

With a better understanding of the basis for the relationship between task similarity and catastrophic forgetting, a natural question to ask is how various commonly used methods to combat forgetting impact the picture. These methods typically fall into three broad groups: dynamic ar-

chitectures, where capacity is added to deal with new tasks; regularisation, where a penalty is added to the objective of later tasks to bias the network to solutions compatible with earlier tasks; and replay, where data from previous tasks (or representations thereof) are interleaved throughout training of later tasks. By using the conventional continual learning protocol of one head per task, we are arguably already operating in the dynamic architecture regime. Beyond that, we study in this section one of the most widely used algorithms for combating forgetting, EWC; as well as interleaved replay, which we can implement straightforwardly in the teacher-student framework without storing data or training additional generative models.

4.1. Elastic Weight Consolidation

EWC (Kirkpatrick et al., 2017) applies a quadratic penalty to weight movement away from the solution for an earlier task and is modulated by the Fisher information of the weight for the earlier task. The penalty is motivated by a Laplace approximation of the posterior (conditional probability of the parameters given the data from the first task) where the mean and variance of the Gaussian approximation are given by the weights at the end of the first task, and the diagonal of the Fisher information matrix respectively. For a pair of tasks A and B, and a neural network parameterised by θ, the objective function for training on the second task is thus given by:

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \frac{\lambda}{2}\sum_i F_i(\theta_i - \theta^*_{A,i})^2, \qquad (4)$$

where F_i is the i^{th} element along the diagonal of the Fisher information matrix, $\theta^*_{A,i}$ is the value of the i^{th} weight at the end of training on task A, and λ is an additional hyperparameter controlling the strength of consolidation. Specifically in the online learning setting of the two-layer student-teacher framework, EWC affects only the first layer weights since the head weights are not shared across tasks.

In Fig. 5 we show the generalisation error curves for a student trained on a succession of two teachers with various degrees of similarity, this time training with the modified EWC objective in the second phase. Each subplot shows a different value for the strength of consolidation (λ in Eq. 4).

As the importance parameter increases and more weight is given to consolidation in the objective, forgetting generally reduces. In particular we see that the more similar the tasks are, the greater λ needs to be to have an impact. Eventually for the largest value of λ shown, all trajectories have collapsed onto a very similar learning trajectory. The exception is the trajectory corresponding to fully aligned teachers, which despite some improvement is comparatively less affected by EWC. We can understand these results through the lens of Maslow's hammer: for fully aligned teachers,

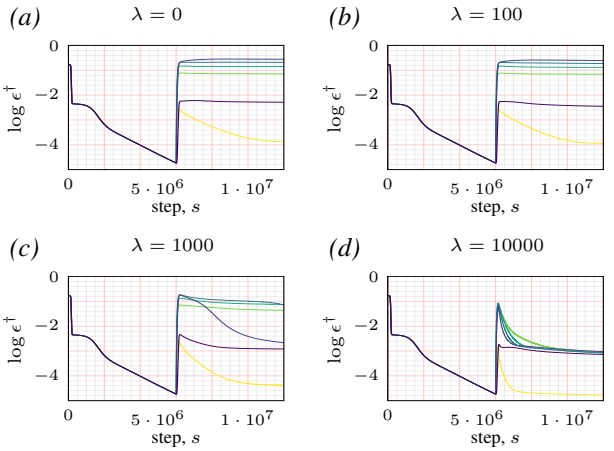


Figure 5. **Effect of EWC on task similarity vs. forgetting** As the consolidation strength λ increases, so too does the bias towards node activation. This happens first for networks trained on more similar teachers until for $\lambda = 10000$, all trajectories regardless of V have essentially collapsed onto one.

the student does not need to activate dormant nodes in the second task and can feasibly continue to use the specialised node, hence EWC has little effect. As the task similarity reduces, the propensity of the student to instead activate a new node increases. The effect of EWC is to intensify this increased propensity, in other words to amplify the bias to fresh node activation since movement in the weights contributing to the specialised node is penalised. Additionally, as the bias towards node re-use is higher for more similar tasks, it takes a stronger push in the other direction (i.e. higher λ) to impact these trajectories. With a high enough λ such that the bias to node re-use is effectively removed for all trajectories regardless of teacher similarity, the task in the second phase of learning is akin to learning a new random teacher with a *tabula rasa* node. Hence the learning trajectories collapse onto one. We show later that although the conservatism induced by strong EWC (comparable in this setting to freezing a node), limits even the fully aligned setting, it can still be a favourable method in intermediate similarity settings.

4.2. Interleaved Replay

One set of methods for combating forgetting involves showing examples from previous tasks during training of later tasks. This can be done in a range of ways from explicitly storing data from previous tasks to training a generative model from which to sample data during later tasks (Shin et al., 2017; Draelos et al., 2017). These methods are inspired by systems consolidation theories in neuroscience, e.g. hippocampal replay (Kumaran et al., 2016).

In the student-teacher framework, it is straightforward to implement this kind of algorithm since the teacher *is* the generative model; this means we can intermittently sample

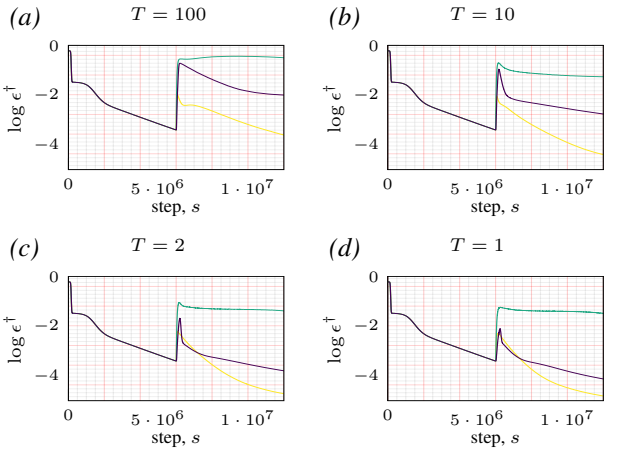


Figure 6. **Effect of replay on task similarity vs. forgetting** Although interleaving is superior to EWC in highly aligned or highly orthogonal regimes, it remains very poor in intermediate regimes even for fully interleaved training.

from previous teachers during training on later teachers. Here we focus in particular on interleaving a single example from the first teacher at different periods, T, during training of the second teacher. Other durations of interleaving and study of potential prioritisation schemes are left for future work. This sensitivity analysis is shown in Fig. 6.

As the period of interleaving reduces, forgetting reduces and indeed the student begins to co-learn in the orthogonal and aligned regimes (this can be thought of as a kind of backward-transfer). Unlike for EWC where high λ collapses the trajectories of the student onto one regardless of teacher-teacher similarity, even for the strongest interleaving ($T = 1$) intermediate similarity remains the most difficult regime. Again this boils down to a difficult and ultimately costly trade-off between node re-use and node activation which interleaved training, unlike strong regularisation, cannot mitigate (see App. H).

It is informative to compare directly the generalisation error trajectories for vanilla SGD, strong interleaving, and strong EWC (see Fig. 7a). Although interleaving is far superior in the orthogonal and aligned regimes (effectively allowing backward transfer), EWC is better in the intermediate similarity case despite heavily stagnating on the first task due to the strong deviation penalty. Given interleaving is generally considered to be the gold standard (Kumaran et al., 2016), this may be unexpected. Is this a consequence of poor initialisation (Liu et al., 2020; Gerace et al., 2022), or is the intermediate similarity task structure inherently challenging for interleaved training? We investigated this question by comparing to the drastic strategy of completely re-initialising at the task boundary and performing interleaved training from a *tabula rasa* network. As we show in Fig. 7b, there is a clear benefit in terms of forgetting in interleaving experiences starting from the solution to the

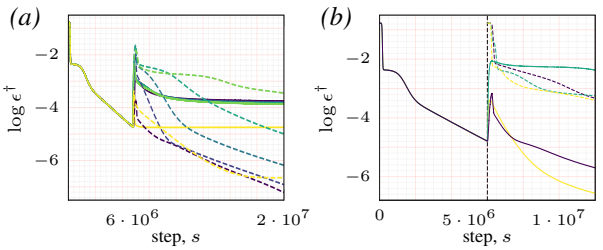


Figure 7. **EWC vs. interleaved replay** *(a)* Generalisation error on first teacher $\log \epsilon^\dagger$ vs step s for a set of teacher-teacher similarities V. Dashed (solid) lines represent $T = 1$ interleaved replay of the first teacher ($\lambda = 1000$ EWC penalty) in the second phase of training. Although interleaving is far superior on the extremes of the spectrum, there is a range of task similarity where interleaved replay is worse than the trajectories of EWC. **Catastrophic slowing** *(b)* Generalisation error $\log \epsilon^\dagger$ vs. step s for extremes of task similarity spectrum V. Solid lines represent interleaved replay from the task switch whereas dashed lines represent a re-initialisation followed by interleaved training from the task switch. If you allow access to the first teacher during the second phase of training, re-initialising at the task switch is better for forgetting (and transfer) in the intermediate task similarity regime.

first task when teachers are orthogonal or aligned. In the intermediate regime however re-initialising entirely is actually better! This cannot be explained away by a trade-off with superior performance on the second task as this is also better when re-initialising (see App. I). This suggests the presence of a *catastrophic slowing* effect in the intermediate regime where interleaving is not a viable combating method due to the tight balance between re-use and activation.

5. Discussion

This work has introduced the Maslow's hammer hypothesis, which shows how a trade-off between re-use and activation at the node level gives rise to a non-monotonic relationship between task similarity and forgetting such that intermediate task similarity is most damaging. The universality of this explanation remains to be fully established: In the teacher-student setup, the isotropy of the input distribution implies that the feature weights must pay attention to every part of the input distribution, and behavioural signatures at the node level capture those of interest at the weight level. In the data-mixing framework, where this data isotropy is broken and we can expect the learned representation in the feature weights to be sparser, we still see evidence for a trade-off between re-use and activation at the node level even if this is a coarsening of more complicated interactions at the weight level. In other settings, this trade-off may play out at a super-node level in clusters or sub-networks. We leave the study of these trade-offs over different components of the network for future work. We use the insights gained from Maslow's hammer to rethink two popular combating methods for catastrophic forgetting (EWC and in-

terleaved replay), and identify among other properties of these methods an effect we term *catastrophic slowing* where interleaving experiences in an intermediate task similarity regime is worse than re-initialising the network, from both a transfer and forgetting perspective. Moving forward, we hope that Maslow's hammer for catastrophic forgetting can help elucidate related phenomena in continual learning and cognate paradigms.

Code: github.com/seblee97/student_teacher_catastrophic

6. Acknowledgements

SL is supported by an EPSRC DTP Studentship. CC would like to acknowledge support from BBSRC (BB/N013956/1, BB/N019008/1), Wellcome Trust (200790/Z/16/Z), the Simons Foundation (564408) and EPSRC(EP/R035806/1). This work was further supported by the Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3755) (SL/SSM/AS). AS is a CIFAR Azrieli Global Scholar in the Learning in Machines & Brains programme.

References

Asanuma, H., Takagi, S., Nagano, Y., Yoshida, Y., Igarashi, Y., and Okada, M. Statistical mechanical analysis of catastrophic forgetting in continual learning with teacher and student networks. *Journal of the Physical Society of Japan*, 90(10):104001, 2021. doi: 10.7566/JPSJ.90.104001. URL https://doi.org/10.7566/JPSJ.90.104001.

Barnett, S. M. and Ceci, S. J. When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological bulletin*, 128(4):612, 2002.

Bell, S. J. and Lawrence, N. D. Behavioral experiments for understanding catastrophic forgetting. *arXiv preprint arXiv:2110.10570*, 2021.

Calvert, G., Spence, C., Stein, B. E., et al. *The handbook of multisensory processes*. MIT press, 2004.

Cichon, J. and Gan, W.-B. Branch-specific dendritic ca 2+ spikes cause persistent synaptic plasticity. *Nature*, 520 (7546):180–185, 2015.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019. URL http://arxiv.org/abs/1810.04805.

Dhifallah, O. and Lu, Y. M. Phase transitions in transfer learning for high-dimensional perceptrons. *Entropy*, 23 (4), 2021. doi: 10.3390/e23040400. URL https://www.mdpi.com/1099-4300/23/4/400.

Doan, T., Bennani, M. A., Mazoure, B., Rabusseau, G., and Alquier, P. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *International Conference on Artificial Intelligence and Statistics*, pp. 1072–1080. PMLR, 2021.

Draelos, T. J., Miner, N. E., Lamb, C. C., Cox, J. A., Vineyard, C. M., Carlson, K. D., Severa, W. M., James, C. D., and Aimone, J. B. Neurogenesis deep learning: Extending deep networks to accommodate new classes. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 526–533. IEEE, 2017.

Engel, A. and Van den Broeck, C. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Flesch, T., Balaguer, J., Dekker, R., Nili, H., and Summerfield, C. Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, 115(44):E10313–E10322, 2018.

Gardner, E. and Derrida, B. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983–1994, 1989.

Gepperth, A. and Karaoguz, C. A bio-inspired incremental learning architecture for applied perceptual problems. *Cognitive Computation*, 8(5):924–934, 2016.

Gerace, F., Saglietti, L., Sarao Mannelli, S., Saxe, A., and Zdeborová, L. Probing transfer learning with a model of synthetic correlated datasets. *Machine Learning: Science and Technology*, 2022. URL http://iopscience.iop.org/article/10.1088/2632-2153/ac4f3f.

Goldt, S., Advani, M., Saxe, A. M., Krzakala, F., and Zdeborová, L. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems*, pp. 6979–6989, 2019.

Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys. Rev. X*, 10:041044, Dec 2020. doi: 10.1103/PhysRevX.10.041044. URL https://link.aps.org/doi/10.1103/PhysRevX.10.041044.

Goodfellow, I. J., Mirza, M., Da, X., Courville, A. C., and Bengio, Y. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB,*

Canada, April 14-16, 2014, Conference Track Proceedings, 2014. URL http://arxiv.org/abs/1312.6211.

Hinton, G. E. and Plaut, D. C. Using fast weights to deblur old memories. In *Proceedings of the 9th Annual Conference of the Cognitive Science Society*, pp. 177–186, 1987.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 32*, pp. 8571–8580, 2018.

Kemker, R., McClure, M., Abitino, A., Hayes, T., and Kanan, C. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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

Kumaran, D., Hassabis, D., and McClelland, J. L. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016.

Lee, S., Goldt, S., and Saxe, A. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pp. 6109–6119. PMLR, 2021.

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

Liu, S., Papailiopoulos, D., and Achlioptas, D. Bad global minima exist and sgd can reach them. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8543–8552. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/618491e20a9b686b79e158c293ab4f91-Paper.pdf.

Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M., Thomas, M. S., and Westermann, G. *Neuroconstructivism-I: How the brain constructs cognition*. Oxford University Press, 2007.

Maslow, A. H. The psychology of science a reconnaissance. 1966.

McClelland, J., McNaughton, B., and O'Reilly, R. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419–57, July 1995.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Mendez, J. A. and EATON, E. Lifelong learning of compositional structures. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=ADWd4TJO13G.

Mézard, M., Parisi, G., and Virasoro, M. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Mirzadeh, S. I., Chaudhry, A., Hu, H., Pascanu, R., Gorur, D., and Farajtabar, M. Wide neural networks forget less catastrophically. *arXiv preprint arXiv:2110.11526*, 2021.

Ostapenko, O., Rodriguez, P., Caccia, M., and Charlin, L. Continual learning via local module composition. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=LJjC6DmSkgT.

Pallier, C., Dehaene, S., Poline, J.-B., LeBihan, D., Argenti, A.-M., Dupoux, E., and Mehler, J. Brain imaging of language plasticity in adopted adults: Can a second language replace the first? *Cerebral cortex*, 13(2):155–161, 2003.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks*, 2019.

Rajkomar, A., Dean, J., and Kohane, I. Machine learning in medicine. *New England Journal of Medicine*, 380(14): 1347–1358, 2019.

Ramasesh, V. V., Dyer, E., and Raghu, M. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=LhY8QdUGSuw.

Ratcliff, R. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. icarl: Incremental classifier and representation learning.

In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Refinetti, M., Goldt, S., Krzakala, F., and Zdeborova, L. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8936–8947. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/refinetti21b.html.

Riegler, P. and Biehl, M. On-line backpropagation in two-layered neural networks. *Journal of Physics A: Mathematical and General*, 28(20), 1995.

Robins, A. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Saad, D. *On-line learning in neural networks*, volume 17. Cambridge University Press, 2009.

Saad, D. and Solla, S. A. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, 74(21):4337, 1995.

Saglietti, L., Mannelli, S. S., and Saxe, A. An analytical theory of curriculum learning in teacher-student networks. *arXiv preprint arXiv:2106.08068*, 2021.

Seung, H. S., Sompolinsky, H., and Tishby, N. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

Shen, Y., Dasgupta, S., and Navlakha, S. Algorithmic insights on continual learning from fruit flies. *arXiv preprint arXiv:2107.07617*, 2021.

Shin, H., Lee, J. K., Kim, J., and Kim, J. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, pp. 2990–2999, 2017.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. A survey on deep transfer learning. In *International conference on artificial neural networks*, pp. 270–279. Springer, 2018.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J., and Hassabis, D. Highly accurate protein structure prediction for the human proteome. *Nature*, 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03828-1. URL https://doi.org/10.1038/s41586-021-03828-1.

Veniat, T., Denoyer, L., and Ranzato, M. Efficient continual learning with modular networks and task-driven priors. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=EKV158tSfwv.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017.

Yang, G., Lai, C. S. W., Cichon, J., Ma, L., Li, W., and Gan, W.-B. Sleep promotes branch-specific formation of dendritic spines after learning. *Science*, 344(6188):1173–1178, 2014.

Yoshida, Y. and Okada, M. Data-dependence of plateau phenomenon in learning with neural network — statistical mechanical analysis. In *Advances in Neural Information Processing Systems 32*, pp. 1720–1728, 2019.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Zhou, G., Sohn, K., and Lee, H. Online incremental feature learning with denoising autoencoders. In *Artificial intelligence and statistics*, pp. 1453–1461. PMLR, 2012.

Appendix

A. Experiment Details

Unless mentioned otherwise in the main text, the following parameters were used in all teacher-student runs:

- Input dimension = 1000;
- 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: 4;
- Teacher hidden dimension: 2;
- Learning rate: 0.1;
- Nonlinearity = scaled error function.

For the data mixing framework, we use the Fashion Modified National Institute of Standards and Technology (MNIST) dataset, with the following parameters:

- $\mathcal{D}_1$: class 0, class 5;
- $\tilde{\mathcal{D}}_2$: class 2, class 7;
- SGD optimiser;
- Mean squared error loss;
- Batch size = 1;
- Input dimension = 1024;
- Hidden dimension = 8;
- Nonlinearity = sigmoid;
- Learning rate: 0.001.

We grayscale the data and apply an early stopping regime such that the weights used for the network in the second task are those with the lowest test error obtained during the first phase of training. This is to avoid any additional effects from overfitting.

Code for the experiments can be found at: github.com/seblee97/student_teacher_catastrophic

B. Specialisation Assumption

Implicit in some of the discussion around Maslow's hammer is that there is specialisation in the network during the first task, and that there is additional capacity in the network to learn later tasks. We show in the ODE solutions for the teacher-student setup and empirically in the data mixing framework that this is the case (even in a relatively small network). We do not show the plots here, but we also confirmed this assumption to be robust in the teacher-student framework over every aspect of the problem we looked at (activation function, level of over-parameterisation, size of hidden dimension, noise and noiseless teachers). In settings where this assumption does not hold at all, Maslow's hammer will likely not hold in the way it is currently stated. We leave investigation of this regime for future work.

C. ODE Formulation

Below we outline the ODE analysis that is used in the teacher-student components of this paper. Since our setup is very similar to that of Lee et al. (2021), the derivations below are also the same. We reproduce it here to facilitate a self-contained paper.

C.1. Order Parameters

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

$$\text{Student-Student Overlap, } \mathbf{Q}: q_{kl} \equiv \langle \lambda_k \lambda_l \rangle = \frac{1}{N} \mathbf{w}_k \mathbf{w}_l; \tag{5}$$

$$\text{Teacher}^\dagger\text{-Teacher}^\dagger\text{Overlap, } \mathbf{T}: t_{nm} \equiv \langle \rho_m \rho_n \rangle = \frac{1}{N} \mathbf{w}_m^\dagger \mathbf{w}_n^\dagger; \tag{6}$$

$$\text{Student-Teacher}^\dagger\text{Overlap, } \mathbf{R}: r_{km} \equiv \langle \lambda_k \rho_m \rangle = \frac{1}{N} \mathbf{w}_k \mathbf{w}_m^\dagger; \tag{7}$$

$$\text{Teacher}^\ddagger\text{-Teacher}^\ddagger\text{Overlap, } \mathbf{S}: s_{pq} \equiv \langle \eta_p \eta_q \rangle = \frac{1}{N} \mathbf{w}_p^\ddagger \mathbf{w}_q^\ddagger; \tag{8}$$

$$\text{Student-Teacher}^\ddagger\text{Overlap, } \mathbf{U}: u_{kp} \equiv \langle \lambda_k \eta_p \rangle = \frac{1}{N} \mathbf{w}_k \mathbf{w}_p^\ddagger; \tag{9}$$

$$\text{Teacher}^\dagger\text{-Teacher}^\ddagger\text{Overlap, } \mathbf{V}: v_{mp} \equiv \langle \rho_m \eta_p \rangle = \frac{1}{N} \mathbf{w}_m^\dagger \mathbf{w}_p^\ddagger; \tag{10}$$

along with the head weights, $\mathbf{v}^\dagger$, $\mathbf{v}^\ddagger$, $\mathbf{h}^\dagger$, $\mathbf{h}^\ddagger$.

Throughout, we denote any quantities associated with the first task with †, any quantity associated with the second task with ‡. In any quantity or equation that generally holds for † or ‡, we represent this by marking it with $*$.

C.2. Generalisation Error in terms of Order Parameters

Our aim is to formulate the generalisation error in terms of the macroscopic order parameters.

The SGD update equations in the two-layer teacher-student setup are given by:

$$\mathbf{w}_k^{\mu+1} = \mathbf{w}_k^\mu - \frac{\alpha_\mathbf{W}}{\sqrt{D}} v_k^{*\mu} g'(\lambda_k^\mu) \Delta^{*\mu} \mathbf{x}^\mu \tag{11a}$$

$$h_k^{*\mu+1} = h_k^{*\mu} - \frac{\alpha_\mathbf{h}}{D} g(\lambda_k^\mu) \Delta^{*\mu}, \tag{11b}$$

where $\alpha_\mathbf{W}$ is the learning rate for the feature weights, $\alpha_\mathbf{h}$ is the learning rate for the head weights, and

$$\Delta^{\dagger\mu} \equiv \sum_k h_k^{\dagger\mu} g(\lambda_k^\mu) - \sum_m v_m^\dagger g(\rho_m^\mu); \tag{12}$$

$$\Delta^{\ddagger\mu} \equiv \sum_k h_k^{\ddagger\mu} g(\lambda_k^\mu) - \sum_p v_p^\ddagger g(\eta_p^\mu). \tag{13}$$

We have also introduced the *local fields*

$$\rho_m \equiv \frac{\mathbf{w}_m \mathbf{x}}{\sqrt{D}}, \qquad \eta_p \equiv \frac{\mathbf{w}_p \mathbf{x}}{\sqrt{D}}, \qquad \lambda_k \equiv \frac{\mathbf{w}_k \mathbf{x}}{\sqrt{D}} \tag{14}$$

of the m^{th} teacher † unit, n^{th} teacher ‡ unit, and k^{th} student unit, respectively. In general, indices i, j, k, l are used for hidden units of the student; m, n for hidden units of †; and p, q for hidden units of ‡.

Let us begin by multiplying out Eq. 1,

$$\epsilon_g^\dagger = \frac{1}{2} \left\langle \left[\sum_{i,k} h_i^\dagger h_k^\dagger g(\lambda_i) g(\lambda_k) + \sum_{m,n} v_m^\dagger v_n^\dagger g(\rho_m) g(\rho_n) - 2 \sum_{i,n} h_i^\dagger v_n^\dagger g(\lambda_i) g(\rho_n) \right] \right\rangle. \tag{15}$$

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

$$\mathcal{P}(\beta,\gamma) = \frac{1}{\sqrt{(2\pi)^{F+H}|\tilde{\mathbf{C}}|}} \exp\left\{-\frac{1}{2}(\beta,\gamma)^T \tilde{\mathbf{C}}^{-1}(\beta,\gamma)\right\}, \tag{16}$$

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

$$I_2(f,h) \equiv \langle g(\beta)g(\gamma)\rangle, \tag{17}$$

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

$$\epsilon_g^\dagger = \frac{1}{2}\sum_{i,k} h_i^\dagger h_k^\dagger I_2(i,k) + \frac{1}{2}\sum_{n,m} v_n^\dagger v_m^\dagger I_2(n,m) - \sum_{i,n} h_i^\dagger v_n^\dagger I_2(i,n) \tag{18}$$

$$\epsilon_g^\ddagger = \frac{1}{2}\sum_{i,k} h_i^\ddagger h_k^\ddagger I_2(i,k) + \frac{1}{2}\sum_{p,q} v_p^\ddagger v_q^\ddagger I_2(p,q) - \sum_{i,p} h_i^\ddagger v_p^\ddagger I_2(i,p). \tag{19}$$

C.2.1. 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, 1995):

$$I_2(i,k) = \frac{1}{\pi}\arcsin\frac{q_{ik}}{\sqrt{(1+q_{ii})(1+q_{kk})}}. \tag{20}$$

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

$$\epsilon_g^\dagger = \frac{1}{\pi}\sum_{i,k} h_i^\dagger h_k^\dagger \arcsin\frac{q_{ik}}{\sqrt{(1+q_{ii})(1+q_{kk})}} + \frac{1}{\pi}\sum_{n,m} v_n^\dagger v_m^\dagger \arcsin\frac{t_{nm}}{\sqrt{(1+t_{nn})(1+t_{mm})}} + \frac{2}{\pi}\sum_{i,n} h_i^\dagger v_n^\dagger \arcsin\frac{r_{in}}{\sqrt{(1+q_{ii})(1+t_{nn})}} \tag{21}$$

$$\epsilon_g^\ddagger = \frac{1}{\pi}\sum_{i,k} h_i^\ddagger h_k^\ddagger \arcsin\frac{q_{ik}}{\sqrt{(1+q_{ii})(1+q_{kk})}} + \frac{1}{\pi}\sum_{p,q} v_p^\ddagger v_q^\ddagger \arcsin\frac{s_{pq}}{\sqrt{(1+s_{pp})(1+s_{qq})}} + \frac{2}{\pi}\sum_{i,p} h_i^\ddagger v_p^\ddagger \arcsin\frac{u_{ip}}{\sqrt{(1+q_{ii})(1+s_{pp})}}. \tag{22}$$

C.3. 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. 11a & Eq. 11b). 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. 11b instead.

C.3.1. ODE FOR $\mathbf{R}$

Consider the inner product of Eq. 11a (in the case of * = †) with $\mathbf{w}_n^\dagger$:

$$\mathbf{w}_k^{\mu+1}\mathbf{w}_n^\dagger - \mathbf{w}_k^\mu \mathbf{w}_n^\dagger = -\frac{\alpha_\mathbf{W}}{\sqrt{D}} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\mathbf{x}^\mu \mathbf{w}_n^\dagger \tag{23}$$

$$= -\alpha_\mathbf{W} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\rho_n^\mu \tag{24}$$

$$r_{kn}^{\mu+1} - r_{kn}^\mu = -\frac{\alpha_\mathbf{W}}{D} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\rho_n^\mu \tag{25}$$

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:

$$\frac{dr_{in}}{d\tau} = -\alpha_\mathbf{W} h_i^\dagger \langle g'(\lambda_i)\Delta^\dagger \rho_n\rangle, \tag{26}$$

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

C.3.2. ODE FOR $\mathbf{Q}$

Consider squaring Eq. 11a (here we can simply use * to denote training on either teacher).

$$\mathbf{w}_k^{\mu+1}\mathbf{w}_i^{\mu+1} - \mathbf{w}_k^\mu \mathbf{w}_i^\mu = -\frac{\alpha_\mathbf{W}}{\sqrt{D}} h_i^{*\mu} g'(\lambda_i^\mu)\Delta^{*\mu}\mathbf{x}^\mu \mathbf{w}_k^\mu - \frac{\alpha_\mathbf{W}}{\sqrt{D}} h_k^{*\mu} g'(\lambda_k^\mu)\Delta^{*\mu}\mathbf{x}^\mu \mathbf{w}_i^\mu$$
$$+ \frac{\alpha_\mathbf{W}^2}{D} h_i^{*\mu} g'(\lambda_i^\mu) h_k^{*\mu} g'(\lambda_k^\mu)(\Delta^{*\mu}\mathbf{x}^\mu)^2 \tag{27}$$

$$= -\alpha_\mathbf{W} h_i^{*\mu} g'(\lambda_i^\mu)\Delta^{*\mu}\lambda_k^\mu - \alpha_\mathbf{W} h_k^{*\mu} g'(\lambda_k^\mu)\Delta^{*\mu}\lambda_i^\mu$$
$$+ \frac{\alpha_\mathbf{W}^2}{D} h_i^{*\mu} g'(\lambda_i^\mu) h_k^{*\mu} g'(\lambda_k^\mu)(\Delta^{*\mu}\mathbf{x}^\mu)^2 \tag{28}$$

$$q_{ki}^{\mu+1} - q_{ki}^\mu = -\frac{\alpha_\mathbf{W}}{D} h_i^{*\mu} g'(\lambda_i^\mu)\Delta^{*\mu}\lambda_k^\mu - \frac{\alpha_\mathbf{W}}{D} h_k^{*\mu} g'(\lambda_k^\mu)\Delta^{*\mu}\lambda_i^\mu$$
$$+ \frac{\alpha_\mathbf{W}^2}{D^2} h_i^{*\mu} g'(\lambda_i^\mu) h_k^{*\mu} g'(\lambda_k^\mu)(\Delta^{*\mu}\mathbf{x}^\mu)^2. \tag{29}$$

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

$$\frac{dq_{ik}}{d\tau} = -\alpha_\mathbf{W} h_i^* \langle g'(\lambda_i)\Delta^*\lambda_k\rangle - \alpha_\mathbf{W} h_k^* \langle g'(\lambda_k)\Delta^*\lambda_i\rangle + \alpha_\mathbf{W}^2 h_i^* h_k^* \langle g'(\lambda_i) g'(\lambda_k)\Delta^{*2}\rangle. \tag{30}$$

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.3.3. ODE FOR $\mathbf{U}$

Consider the inner product of Eq. 11a (in the case of * = †) with $\mathbf{w}_p^\ddagger$:

$$\mathbf{w}_k^{\mu+1}\mathbf{w}_p^\ddagger - \mathbf{w}_k^\mu \mathbf{w}_p^\ddagger = -\frac{\alpha_\mathbf{W}}{\sqrt{D}} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\mathbf{x}^\mu \mathbf{w}_p^\ddagger \tag{31}$$

$$= -\alpha_\mathbf{W} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\eta_p^\mu \tag{32}$$

$$u_{kp}^{\mu+1} - u_{kp}^\mu = -\frac{\alpha_\mathbf{W}}{D} h_k^{\dagger\mu} g'(\lambda_k^\mu)\Delta^{\dagger\mu}\eta_p^\mu. \tag{33}$$

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

$$\frac{du_{ip}}{d\tau} = -\alpha_\mathbf{W} h_i^* \langle g'(\lambda_i)\Delta^*\eta_p\rangle. \tag{34}$$

C.3.4. ODE FOR $\mathbf{h}^*$

Here, we simply take the thermodynamic limit of Eq. 11b (for * = †):

$$\frac{dh_i^\dagger}{d\tau} = -\alpha_h \langle \Delta^\dagger g(\lambda_i)\rangle \tag{35}$$

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

$$\Delta^{\dagger\mu} \equiv \sum_k h_k^{\dagger\mu} g(\lambda_k^\mu) - \sum_m v_m^\dagger g(\rho_m^\mu). \tag{36}$$

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

$$\frac{dr_{in}}{d\tau} = -\alpha_{\mathbf{W}} h_i^\dagger \left\langle g'(\lambda_i) \left[\sum_k h_k^\dagger g(\lambda_k) - \sum_m v_m^\dagger g(\rho_m) \right] \rho_n \right\rangle; \tag{37}$$

$$\begin{aligned} \frac{dq_{ik}}{d\tau} = &-\alpha_{\mathbf{W}} h_i^\dagger \left\langle g'(\lambda_i) \left[\sum_j h_j^\dagger g(\lambda_j) - \sum_m v_m^\dagger g(\rho_m) \right] \lambda_k \right\rangle \\ &- \alpha_{\mathbf{W}} h_k^\dagger \left\langle g'(\lambda_k) \left[\sum_j h_j^\dagger g(\lambda_j) - \sum_m v_m^\dagger g(\rho_m) \right] \lambda_i \right\rangle \\ &+ \alpha_{\mathbf{W}}^2 h_i^\dagger h_k^\dagger \left\langle g'(\lambda_i) g'(\lambda_k) \left[\sum_j h_j^\dagger g(\lambda_j) - \sum_m v_m^\dagger g(\rho_m) \right]^2 \right\rangle; \end{aligned} \tag{38}$$

$$\frac{du_{ip}}{d\tau} = -\alpha_{\mathbf{W}} h_i^\dagger \left\langle g'(\lambda_i) \left[\sum_k h_k^\dagger g(\lambda_k) - \sum_m v_m^\dagger g(\rho_m) \right] \eta_p \right\rangle; \tag{39}$$

$$\frac{dh_i^\dagger}{d\tau} = -\alpha_{\mathbf{h}} \left\langle \left[\sum_k h_k^\dagger g(\lambda_k) - \sum_m v_m^\dagger g(\rho_m) \right] g(\lambda_i) \right\rangle. \tag{40}$$

Similarly to the I_2 integral defined in Eq. 17, we further define:

$$I_3(d, f, h) = \langle g'(\zeta) \beta g(\gamma) \rangle, \tag{41}$$

$$I_4(d, e, f, h) = \langle g'(\zeta) g'(\iota) g(\beta) g(\gamma) \rangle; \tag{42}$$

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:

$$\frac{dr_{in}}{d\tau} = \alpha_{\mathbf{W}} h_i^\dagger \left[\sum_m^M v_m^* I_3(i, n, m) - \sum_k^K h_k^\dagger I_3(i, n, k) \right]; \tag{43}$$

$$\begin{aligned} \frac{dq_{ik}}{d\tau} = &\alpha_{\mathbf{W}} h_i^\dagger \left[\sum_m^M v_m^\dagger I_3(i, k, m) - \sum_j^K h_j^\dagger I_3(i, k, j) \right] \\ &+ \alpha_{\mathbf{W}} h_k^\dagger \left[\sum_m^M v_m^\dagger I_3(k, i, m) - \sum_j^K h_j^\dagger I_3(k, i, j) \right] \\ &+ \alpha_{\mathbf{W}}^2 h_i^\dagger h_k^\dagger \left[\sum_{j,l}^K h_j^\dagger h_l^\dagger I_4(i, k, j, l) + \sum_{m,n}^M v_m^\dagger v_n^\dagger I_4(i, k, m, n) \right. \\ &\left. -2 \sum_j^K \sum_m^M v_m^\dagger h_j^\dagger I_4(i, k, j, m) \right]; \end{aligned} \tag{44}$$

$$\frac{du_{ip}}{d\tau} = \alpha_{\mathbf{W}} h_i^\dagger \left[\sum_m^M v_m^\dagger I_3(i, p, m) - \sum_k^K h_k^\dagger I_3(i, p, k) \right]; \tag{45}$$

$$\frac{dh_i^\dagger}{d\tau} = \alpha_{\mathbf{h}} \left[\sum_m^M v_m^\dagger I_2(m,i) - \sum_k^K h_k^\dagger I_2(k,i) \right]. \tag{46}$$

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 Sec. D.1. The next chapter introduces the experimental framework that compliments the theoretical formalism presented above.

D.1. Gaussian Integrals under Scaled Error Function

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

$$I_2(f,h) \equiv \langle g(\beta) g(\gamma) \rangle, \tag{47}$$
$$I_3(d,f,h) \equiv \langle g'(\zeta) \beta g(\gamma) \rangle, \tag{48}$$
$$I_4(d,e,f,h) \equiv \langle g'(\zeta) g'(\iota) g(\beta) g(\gamma) \rangle; \tag{49}$$

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:

$$I_2 = \frac{1}{\pi} \arcsin \frac{c_{12}}{\sqrt{(1+c_{11})(1+c_{22})}}; \tag{50}$$
$$I_3 = \frac{2c_{23}(1+c_{11}) - 2c_{12}c_{13}}{\sqrt{\Lambda_3}(1+c_{11})}; \tag{51}$$
$$I_4 = \frac{4}{\pi^2 \sqrt{\Lambda_4}} \arcsin \frac{\Lambda_0}{\sqrt{\Lambda_1 \Lambda_2}}; \tag{52}$$

where

$$\Lambda_0 = \Lambda_4 c_{34} - c_{23}c_{24}(1+c_{11}) - c_{13}c_{14}(1+c_{22}) + c_{12}c_{13}c_{24} + c_{12}c_{14}c_{23}; \tag{53}$$
$$\Lambda_1 = \Lambda_4(1+c_{33}) - c_{23}^2(1+c_{11}) - c_{13}^2(1+c_{22}) + 2c_{12}c_{13}c_{23}; \tag{54}$$
$$\Lambda_2 = \Lambda_4(1+c_{44}) - c_{24}^2(1+c_{11}) - c_{14}^2(1+c_{22}) + 2c_{12}c_{14}c_{24}; \tag{55}$$
$$\Lambda_3 = (1+c_{11})(1+c_{33}) - c_{13}^2; \tag{56}$$
$$\tag{57}$$

and where c is the relevant projected down covariance matrix.

E. Overlap Generation

Throughout this work we tune the similarity of tasks in the teacher-student setup by manipulating the feature weights of the second teacher in relation to the first. Here we outline the procedure used. Once again this is largely similar to the procedure used in (Lee et al., 2021).

In the main text we mention that we abbreviate the matrix elements v_{mp} to V. The motivation for this is that in the single hidden unit case the input-hidden weights are vectors and their dot products are simply scalars. When we move to the multi-hidden unit setting (see below), there is no single similarity measure from the overlap matrix but we construct our similarity based on a scalar interpolation between two random matrices, so continue to denote similarity simply by V.

For teachers with a single hidden unit we simply need a procedure to generate two D-dimensional vectors (where D is the input dimension), $\mathbf{v}_1$, $\mathbf{v}_2$, with an angle θ between them such that:

$$\mathbf{v}_1 \cdot \mathbf{v}_2 = \theta. \tag{58}$$

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 $D \times D$ orthogonal matrix, R. There is a standard sicpy implementation for this based on QR decomposition of a random Gaussian matrix[1].

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

$$\mathbf{v}_1 = R[:, 1:2] \cdot \tilde{\mathbf{v}}_1; \tag{59}$$
$$\mathbf{v}_2 = R[:, 1:2] \cdot \tilde{\mathbf{v}}_2. \tag{60}$$

For the more general case of multi-hidden units, V is closer to an interpolation than a rotation. Specifically it is an interpolation between two random matrices such that $V = 0$ gives a new random matrix that is orthogonal to the first, and $V = 1$ gives back the same matrix as the first. Formally for similarity measure V and first teacher feature weight matrix, $\mathbf{W}^\dagger$, the second teacher feature weight matrix is given by:

$$\mathbf{W}^\ddagger = V\mathbf{W}^\dagger + \sqrt{(1 - V^2)}\mathbf{Z}, \tag{61}$$

where $\mathbf{Z}$ is a $D \times D$ random matrix.

F. Empirical Node Importance

While specialisation measures are very clearly identifiable via the overlap matrices in the teacher-student, analogues do not exist in the standard supervised learning setup. For this we use an empirical measure of node 'importance' defined as the drop in test error when the node is masked. Formally for a two-layer network, node index i and task index t:

$$I_i^t \equiv \frac{1}{2}\left\langle \left[\sum_{\substack{l=1 \\ l\neq i}}^{L} \mathbf{v}_l g(\mathbf{W}_l \mathbf{x}_j^t) - y_j^t\right]^2 - \left[\sum_{l=1}^{L} \mathbf{v}_l g(\mathbf{W}_l \mathbf{x}_j^t) - y_j^t\right]^2 \right\rangle \tag{62}$$

where $\mathbf{v}$ are the second layer weights, g is the activation function, $\mathbf{W}$ are the first layer weights, and the averages are taken over the test dataset pairs $(\mathbf{x}_j, y_j)$.

[1] SciPy Stats Module Docs

G. Detailed Statistics of Node Re-Use Tendency in Data Mixing

In Fig. 4c, we show the $I^1 \cdot I^2$ vs. α. Specifically we plot the mean of 200 random seeds. At each value of α there is a high level of variance in $I^1 \cdot I^2$. However each decile of the distribution of $I^2 \cdot I^2$ follows a similar monotonic distribution, which we show in the plots below. This demonstrates that while there is a large width in the distribution of $I^1 \cdot I^2$, there is a systematic shift upwards as α increases from 0 to 1.

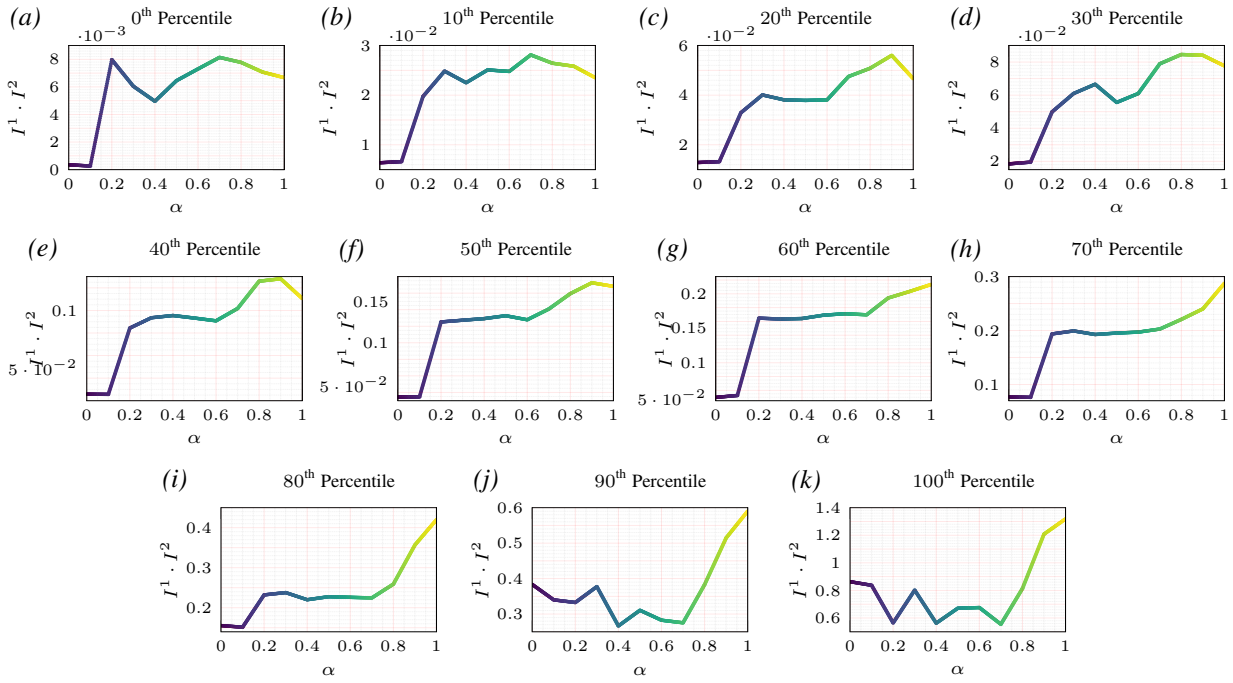


Figure 8. **Deciles of $I^1 \cdot I^2$ distributions over 200 seeds vs. mixing parameter α** In every decile there is a general monotonic relationship between $I^1 \cdot I^2$ and the mixing parameter α.

H. Overlap Plots

In this section we show plots of the overlap parameters corresponding to the various figures in the main text. These can help to illuminate some of the behaviours at the node level that give rise to the macroscopic phenomena we observe e.g. in the generalisation errors.

H.1. Effect of EWC on task similarity vs. forgetting

These are the overlaps associated with Fig. 5.

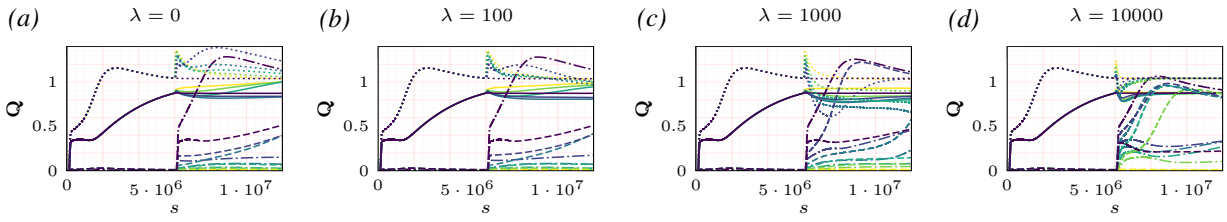


Figure 9. **Effect of EWC on task similarity vs. forgetting: student self-overlaps**

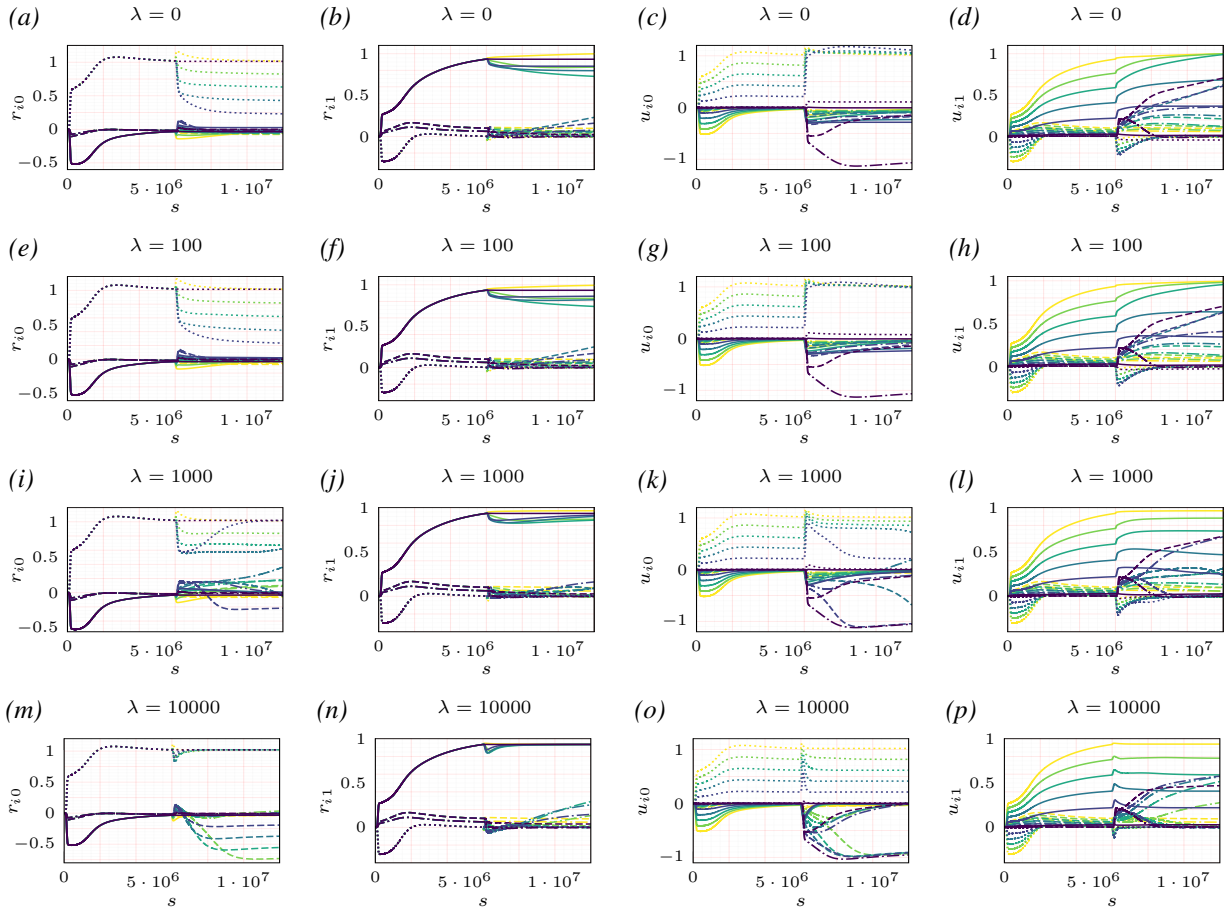


Figure 10. **Effect of EWC on task similarity vs. forgetting: student-teacher overlaps**

H.2. Effect of interleaving on task similarity vs. forgetting

These are the overlaps associated with Fig. 6.

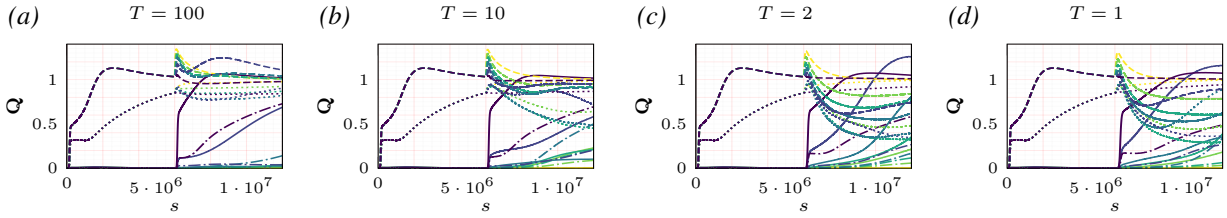


Figure 11. **Effect of interleaving on task similarity vs. forgetting: student self-overlaps**

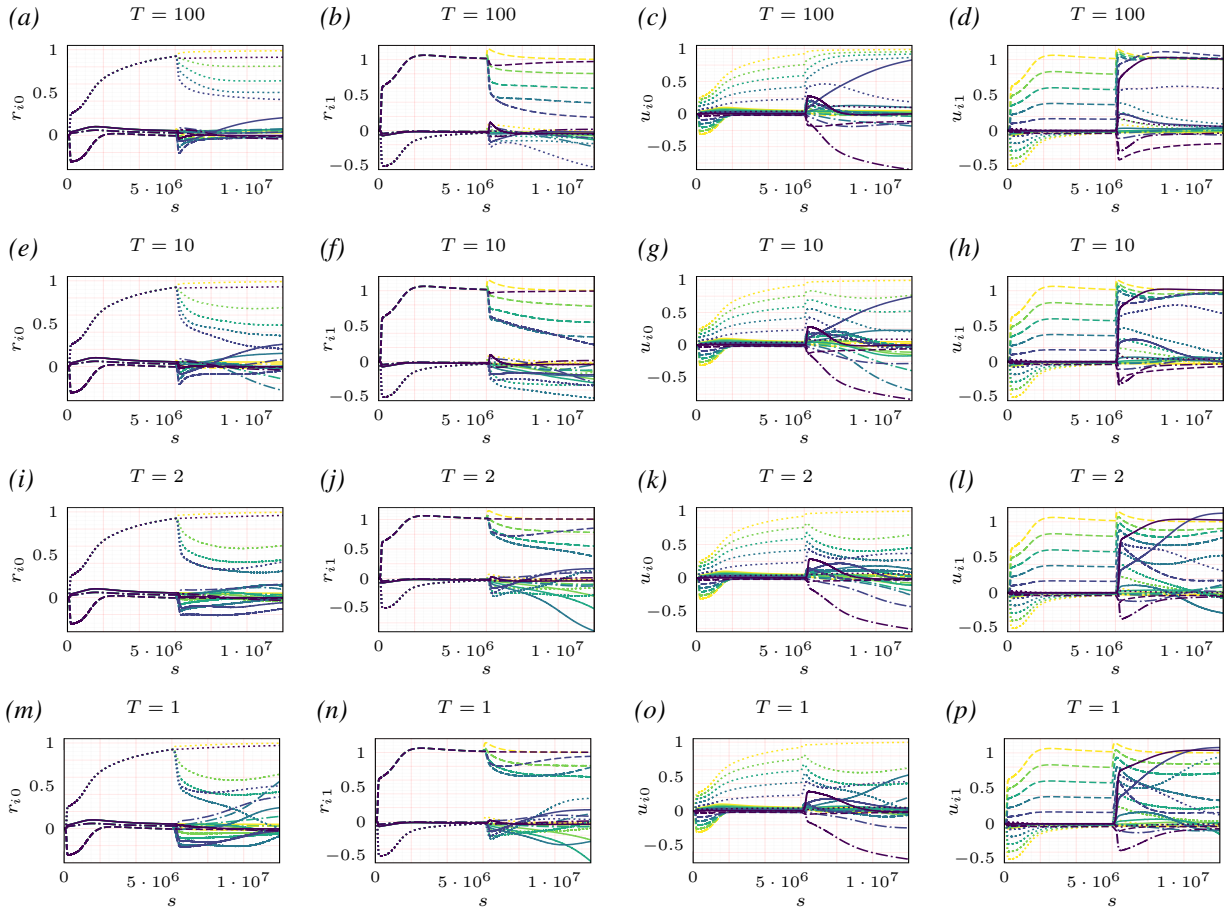


Figure 12. **Effect of interleaving on task similarity vs. forgetting: student-teacher overlaps**

I. Transfer Plots

This section contains plots on the 'transfer' observed in the performance on task 2. We follow (Lee et al., 2021) in defining transfer as the difference between the generalisation error on the second teacher at some step after switch and the generalisation error on the second teacher at the switch. Principally we could ask many of the questions related to forgetting we ask in this paper for transfer as well in the sense that the teacher-student framework permits a natural transfer analogue. However forgetting has been the focus of this work and thus we include transfer plots here for the interested reader.

To help distinguish these plots from those showing quantities associated with forgetting, we use a different color scheme for transfer related plots:

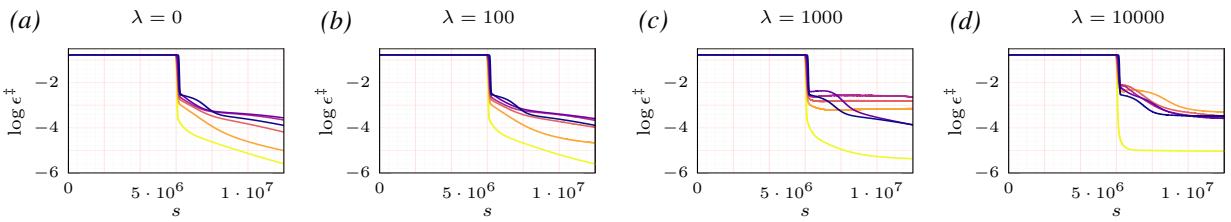


Figure 13. **Effect of EWC on task similarity vs. forgetting: transfer**

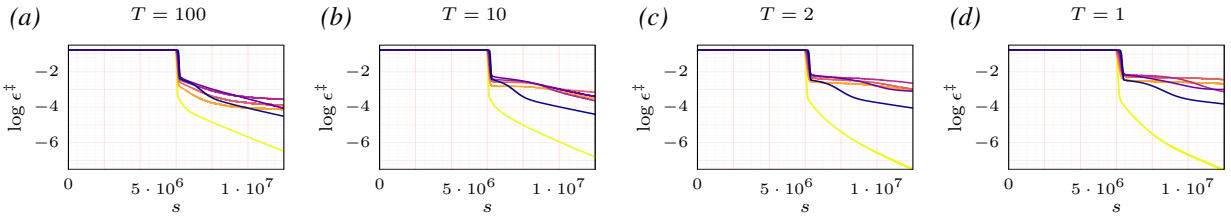


Figure 14. **Effect of replay on task similarity vs. forgetting: transfer**

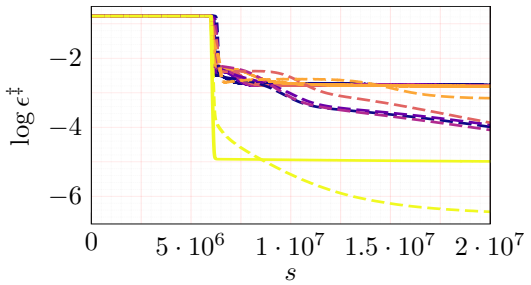


Figure 15. **EWC vs. interleaved replay transfer** In Fig. 7a we compared strong interleaving with strong elastic weight consolidation. Here we show the transfer performance. The solid lines show EWC and the dashed lines show interleaved training. You can see clearly that the performance of EWC plateaus; this is because movement in the node that specialised on the first task is so highly penalised and the network is less flexible as a result. On the other hand transfer performance (as well as backward transfer to the first task as seen by the plot in the main text) continues to improve under interleaved training for highly aligned and orthogonal tasks.

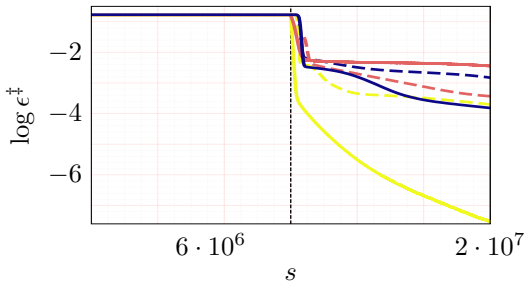


Figure 16. **Catastrophic slowing transfer** In Fig. 7b we identify an effect we call *catastrophic slowing* where interleaved replay can not aid forgetting in the intermediate task regime. Here we show that transfer is also poor in this regime. The dashed lines show trajectories when we re-initialise at the task boundary. While transfer (and forgetting) is better for the aligned and orthogonal cases under interleaving than under re-initialising, it is worse for intermediately related tasks.

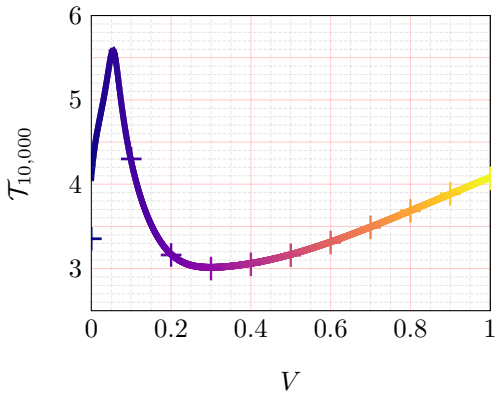


Figure 17. **Transfer vs. task similarity**: This is the transfer equivalent of Fig. 1b. For a full discussion of the implications of this plot, see (Lee et al., 2021) from which this is reproduced.