Continual Learning in the Teacher-Student Setup: Impact of Task Similarity

Sebastian Lee [1] Sebastian Goldt [2] Andrew Saxe [3 4 5]

Abstract

Continual learning—the ability to learn many tasks in sequence—is critical for artificial learning systems. Yet standard training methods for deep networks often suffer from catastrophic forgetting, where learning new tasks erases knowledge of earlier tasks. While catastrophic forgetting labels the problem, the theoretical reasons for interference between tasks remain unclear. Here, we attempt to narrow this gap between theory and practice by studying continual learning in the teacher-student setup. We extend previous analytical work on two-layer networks in the teacher-student setup to multiple teachers. Using each teacher to represent a different task, we investigate how the relationship between teachers affects the amount of forgetting and transfer exhibited by the student when the task switches. In line with recent work, we find that when tasks depend on similar features, intermediate task similarity leads to greatest forgetting. However, feature similarity is only one way in which tasks may be related. The teacher-student approach allows us to disentangle task similarity at the level of *readouts* (hidden-to-output weights) and *features* (input-to-hidden weights). We find a complex interplay between both types of similarity, initial transfer/forgetting rates, maximum transfer/forgetting, and long-term transfer/forgetting. Together, these results help illuminate the diverse factors contributing to catastrophic forgetting.

[1]Imperial College, London, UK [2]International School of Advanced Studies (SISSA), Trieste, Italy [3]Department of Experimental Psychology, University of Oxford, UK [4]CIFAR Azrieli Global Scholars program, CIFAR, Toronto, Canada [5]Facebook AI Research. Correspondence to: Sebastian Lee <sebastian.lee14@imperial.ac.uk>, Andrew Saxe <andrew.saxe@psy.ox.ac.uk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1. Introduction

One of the biggest open challenges in machine learning is the ability to effectively perform continual learning: learning tasks sequentially. A significant hurdle in getting systems to do this effectively is that models trained on task A followed by task B will struggle to learn task B without un-learning task A. This is known as catastrophic interference or *catastrophic forgetting* (McCloskey & Cohen, 1989; Goodfellow et al., 2013), which occurs because weights that contain important information for the first task are overwritten by information relevant to the second. The harmful effects of catastrophic forgetting are not limited to continual learning. They also play a role in multi-task learning, reinforcement learning and standard supervised learning, for example under distribution shift (Arivazhagan et al., 2019; Toneva et al., 2018).

As a result, the phenomenon has received increased interest in recent years. In neuroscience, much work has been done to understand the brain's ability to consolidate learning from earlier tasks, thereby making it relatively robust to forgetting (Flesch et al., 2018; Cichon & Gan, 2015; Yang et al., 2014). Similarly, a series of works has started a systematic empirical analysis of this phenomenon in deep networks (Parisi et al., 2019; Mirzadeh et al., 2020; Neyshabur et al., 2020; Nguyen et al., 2019; Ruder & Plank, 2017). These works consistently observed a counter-intuitive role of the similarity between tasks A and B, with *intermediate* task similarity leading to worst forgetting (Ramasesh et al., 2020; Doan et al., 2020; Nguyen et al., 2019).

The purpose of this work is to tackle continual learning from the complementary perspective of high-dimensional teacher-student models (Gardner & Derrida, 1989; Seung et al., 1992; Biehl & Schwarze, 1993; Zdeborová & Krzakala, 2016). These models are a popular framework for studying machine learning problems in a controlled setting, and have recently seen a surge of interest in attempts to understand generalisation in deep neural networks.

Main contributions

- We analyse continual learning in two-layer neural networks by deriving a closed set of equations which predict the test error of the network trained on a succession of tasks using one-pass (or online) SGD, extending

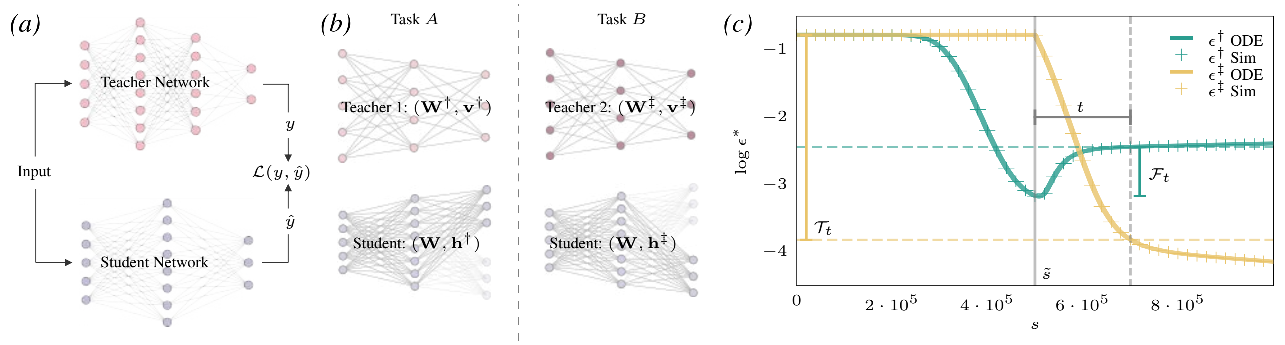

Figure 1. **Continual learning in the teacher-student setup** *(a)* Illustration of the vanilla teacher-student setup, in which a "student" network is trained on i.i.d. inputs with labels from a "teacher" network. *(b)* We model continual learning by training a two-layer *student* (Eqn. 1), on a succession of two teachers, representing distinct tasks A and B. *(c)* Typical generalisation errors of the student (Eq. 2) w.r.t. both teachers during vs. training step s. The solid lines show theoretical predictions derived in subsection 2.1; the crosses are obtained through a single numerical simulation of a networks with input dimension $N = 10^4$. We also label the key quantities of interest in this study: forgetting, $\mathcal{F}_t$ (Eq. 3), and transfer, $\mathcal{T}_t$, (Eq. 4). *Parameters:* $M = 2, K = 1, V = 1$.

classical work on standard supervised learning by Saad & Solla (1995a); Riegler & Biehl (1995).

- Using these equations, we show that intermediate task similarity leads to greatest forgetting in our model.
- We disentangle task similarity on the level of features (input-to-hidden weights) and readouts (input-to-hidden weights) and describe the effect of both types of similarity on forgetting and transfer in infinitely wide networks. We find that feature and readout similarity contribute in complex and sometimes non-symmetric ways to a range of forgetting and transfer metrics.

We summarise our approach in Fig. 1. In the classical teacher-student setup (illustrated in Fig. 1a), a "student" neural network is trained on synthetic data where inputs $\mathbf{x} \in \mathbb{R}^D$ are drawn element-wise i.i.d. from the normal distribution and labels are generated by a "teacher" network (Gardner & Derrida, 1989). To model continual learning, here we consider a setup with two teachers (denoted by $\dagger$ and $\ddagger$), which correspond to two tasks to be learned in succession. Let $\phi(\mathbf{x}; \mathbf{W}, \mathbf{v})$ denote the output of a two-layer network with L hidden neurons, first and second layer weights $\mathbf{W} \in \mathbb{R}^{L\times D}$ and $\mathbf{v} \in \mathbb{R}^L$, and activation g after the hidden layer, *i.e.*

$$\phi(\mathbf{x}; \mathbf{W}, \mathbf{v}) = \sum_{l=1}^{L} \mathbf{v}_l g\left(\frac{\mathbf{W}_l \mathbf{x}}{\sqrt{D}}\right). \tag{1}$$

We generally use K (M) for the number of hidden neurons of the student (teacher). In the first phase of training (left side of Fig. 1b), labels are generated by the first teacher via $y^{\dagger} = \phi(\mathbf{x}; \mathbf{W}^{\dagger}, \mathbf{v}^{\dagger})$, and student outputs are given by $\hat{y}^{\dagger} = \phi(\mathbf{x}; \mathbf{W}, \mathbf{h}^{\dagger})$. Training proceeds using Stochastic Gradient Descent (SGD) on the squared error of $y^{\dagger}$, $\hat{y}^{\dagger}$ in the online regime, where at each step of SGD we draw a new sample (x, y) to evaluate the gradients, until the task *switch*. We follow a standard *multi-headed* approach to continual learning (Zenke et al., 2017; Farquhar & Gal, 2018), in which the student keeps its first-layer weights for the new task, but adds a set of head weights. Thus in the second phase of training, the error is computed over $y^{\ddagger}$, $\hat{y}^{\ddagger}$. Retaining both heads allows us to continually monitor the performance of the student on both tasks after switch, and in theory permits the student to represent both teachers perfectly if given sufficient hidden units.

The generalisation error of the student on the two tasks can be defined as

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

where $*$ denotes either task $\dagger$ or $\ddagger$, and the average $\langle \cdot \rangle$ is taken over the input distribution $\mathbf{x}$ for a given set of teacher and student weights. Note in the online SGD setting, there is no distinction between train and test error. We emphasise that the student has the same set of first-layer weights ($\mathbf{W}$) for both tasks, but different head weights $\mathbf{h}^{\dagger}$, $\mathbf{h}^{\ddagger}$.

Our main theoretical contribution is a set of dynamical equations that predict the evolution of the test error Eq. 2 during the course of training in the limit of large input dimension $D \to \infty$ with $K, M \sim O(1)$, see Sec. 2. We plot the theoretical prediction in Fig. 1c together with a single simulation (crosses); even at moderate input size $D = 10^4$, the agreement is good. We observe that the student error on the first task (green) decreases in the first period of training. After switching tasks at $\tilde{s} = 5 \cdot 10^5$, the error of the student on the second task (yellow) decreases, but the error on the first task increases. We define *forgetting* and *transfer* at time $\tilde{s} + t$ as

$$\text{Forgetting: } \mathcal{F}_t \equiv \epsilon^{\dagger}|_{\tilde{s}+t} - \epsilon^{\dagger}|_{\tilde{s}}, \tag{3}$$

$$\text{Transfer: } \mathcal{T}_t \equiv \epsilon^{\ddagger}|_{\tilde{s}} - \epsilon^{\ddagger}|_{\tilde{s}+t}, \tag{4}$$

see Fig. 1c. An increase in error for the first task after the switch corresponds to positive forgetting, while a reduction in error for the second task corresponds to positive transfer. An alternative definition of transfer would compare the performance of the continual learner on task B to the performance of a student that was trained directly on that task. However, this definition introduces additional hyperparameters which need to be accounted for, such as the distribution of weights at initialisation and at the switching time. Since our focus in this manuscript is on catastrophic forgetting, we focus on the simpler definition of transfer in (4), and leave an exploration of other transfer measures to future work.

A fundamental question in continual learning is the relationship between forgetting/transfer and the task similarity. While one might expect forgetting to decrease with increasing task similarity, Ramasesh et al. (2020)—through a series of careful experiments on the CIFAR10 and CIFAR100 datasets—observed that *intermediate* task similarity leads to greatest forgetting. We were able to reproduce their results for the two-layer neural networks (1), see App. A.The primary objective of this work is now to use our multi-teacher-student setup, which gives us full control over teacher similarity, to analyse dependence of forgetting and transfer on task similarity theoretically.

1.1. Further Related Work

The **teacher-student framework** has a long history in studying the dynamics of learning in neural network models (Gardner & Derrida, 1989; Seung et al., 1992; Watkin et al., 1993; Engel & Van den Broeck, 2001; Zdeborová & Krzakala, 2016) and has recently experienced a surge of activity in the machine learning community (Zimmer et al., 2014; Zhong et al., 2017; Tian, 2017; Du et al., 2018; Soltanolkotabi et al., 2018; Aubin et al., 2018; Saxe et al., 2018; Baity-Jesi et al., 2018; Goldt et al., 2019; Ghorbani et al., 2019; Yoshida & Okada, 2019; Ndirango & Lee, 2019; Gabrié, 2020; Bahri et al., 2020; Zdeborová, 2020; Advani et al., 2020). While this article went to press, a preprint by Asanuma et al. (2021) appeared which analyses continual learning in a teacher-student setup for linear regression.

The teacher-student approach has recently been used to study **transfer learning**, both in linear networks (Lampinen & Ganguli, 2018) and in non-linear perceptron models (Dhifallah & Lu, 2021), which correspond to the $K = M = 1$ case of our setup. While the transfer of knowledge from one task to the next is an important aspect in continual learning, the latter is crucially also interested in the retention–or forgetting–of knowledge about the first task. This can be most clearly seen in the fact that in transfer learning, there is only one set of student head weights. Indeed, we will find an interesting interplay between transfer and forgetting in our models of continual learning.

Continual learning in the NTK regime Doan et al. (2020) analysed the impact of task similarity, and also found increasing task similarity leads to more forgetting. The key difference to our work is that their study focuses on the neural tangent kernel (NTK) (Jacot et al., 2018) or "lazy" regime (Chizat et al., 2019) of two-layer networks, where the first layer of weights stays approximately constant throughout training. Bennani & Sugiyama (2020) gave guarantees on the error achieved with orthogonal gradient descent in the same regime. Here, we focus on the regime where the weights of the network move significantly and are thus able to learn features, which will be key to our analysis in Sec. 2 and to our disentangling of feature vs. readout similarity in Sec. 3.

The **dynamics of two-layer neural networks** trained using online SGD in the classic teacher-student setup of Fig. 1a was first studied in a series of classic papers by Biehl & Schwarze (1995) and Saad & Solla (1995a) who derived a set of closed ODEs that track the test error of the student (see also Saad & Solla (1995b); Biehl et al. (1996); Saad (2009) for further results and Goldt et al. (2019) for a recent proof of these equations). Here, we extend this type of analysis to the continual learning model of Fig. 1b. The aforementioned works all consider the limit of large input dimension $D \to \infty$, while the number of neurons is of order 1. The complementary "mean-field" limit of finite input dimension and an infinite number of hidden neurons was analysed (Mei et al., 2018; Chizat & Bach, 2018; Sirignano & Spiliopoulos, 2020; Rotskoff & Vanden-Eijnden, 2018). We will turn to this limit to disentangle the impact of feature and readout similarity in Sec. 3.

Many methods for **combating catastrophic interference** have been proposed, often taking the form of regularisation, architecture expansion, and/or replay (Parisi et al., 2019; Farquhar & Gal, 2018). Regularisation-based methods constrain weights to retain information about earlier tasks (Zenke et al., 2017; Li & Hoiem, 2017; Kirkpatrick et al., 2017); architectural methods add capacity to the network for each new task (Rusu et al., 2016); and replay methods store data from earlier tasks to interleave when learning new tasks (McClelland et al., 1995; Shin et al., 2017).

2. Continual learning in the large input limit

We begin by studying the impact of task similarity on the dynamics and the performance of learning in the limit of large input dimension $D \to \infty$, while the number of neurons $K, M \sim O(1)$.

Training We train the student using online stochastic gradient descent on the L_2 loss. Each new input $\mathbf{x}$ is fed to the teacher to compute the target output via $y^* = \phi(\mathbf{x}; \mathbf{W}^*, \mathbf{v}^*)$, while the student prediction is given by $\hat{y}^\dagger = \phi(\mathbf{x}; \mathbf{W}, \mathbf{h}^*)$. The student's weights in both layers are updated through gradient descent on $\frac{1}{2}(\hat{y}^* - y^*)^2$. The SGD weight updates are given by:

$$\mathbf{w}_k^{\mu+1} = \mathbf{w}_k^{\mu} - \frac{\alpha_{\mathbf{W}}}{\sqrt{D}} v_k^{*\mu} g'(\lambda_k^\mu) \Delta^{*\mu} \mathbf{x}^\mu \tag{5a}$$

$$h_k^{*\mu+1} = h_k^{*\mu} - \frac{\alpha_{\mathbf{h}}}{D} g(\lambda_k^\mu) \Delta^{*\mu}, \tag{5b}$$

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

$$\Delta^{\dagger\mu} \equiv \sum_k h_k^{\dagger\mu} g(\lambda_k^\mu) - \sum_m v_m^\dagger g(\rho_m^\mu); \tag{6}$$

$$\Delta^{\ddagger\mu} \equiv \sum_k h_k^{\ddagger\mu} g(\lambda_k^\mu) - \sum_p v_p^\ddagger g(\eta_p^\mu). \tag{7}$$

We have also introduced the *local fields*

$$\rho_m \equiv \frac{\mathbf{w}_m \mathbf{x}}{\sqrt{D}}, \qquad \eta_p \equiv \frac{\mathbf{w}_p \mathbf{x}}{\sqrt{D}}, \qquad \lambda_k \equiv \frac{\mathbf{w}_k \mathbf{x}}{\sqrt{D}} \tag{8}$$

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 ‡. Initial weights are taken i.i.d. from the normal distribution with standard deviation σ_0. The different scaling of the learning rates for first and second-layer weights guarantees the existence of a well-defined limit of the SGD dynamics as $D \to \infty$. We make the crucial assumption that at each step of the algorithm, we use a previously unseen sample $(\mathbf{x}, y^*)$. This limit of infinite training data is variously known as online learning or one-shot/single-pass SGD. We note that in general the head weights could also be matrices if a teacher has multiple output nodes, but we focus on the case of a single output here to keep notation light.

The "order parameters" of the problem The key quantity in our analysis is the test error Eq. 2, which (e.g. for †) can be written more explicitly as

$$\epsilon^\dagger(\mathbf{W}, \mathbf{W}^\dagger, \mathbf{h}^\dagger, \mathbf{v}^\dagger) = \frac{1}{2}\left\langle \left[\sum_{k=1}^K h_k^\dagger g(\lambda_k) - \sum_{m=1}^M v_m^\dagger g(\rho_m) \right]^2 \right\rangle. \tag{9}$$

To evaluate the average, the input $\mathbf{x}$ only appears via products with the student weights (λ_k) and likewise for the teacher; we can hence replace the high-dimensional averages over $\mathbf{x}$ with an average over the $K+M$ "local fields" λ and ρ. Since we take the inputs element-wise i.i.d. from the standard Gaussian distribution, we have $\langle x_i \rangle = 0$ and $\langle x_i x_j \rangle = \delta_{ij}$. It also follows immediately that the local fields are jointly Gaussian, with mean $\langle \lambda_k \rangle = \langle \rho_m \rangle = 0$. The test error can hence be written as a function of only the second moments of the joint distribution of (ρ, λ), which we define as

$$q_{kl} \equiv \langle \lambda_k \lambda_l \rangle, \quad r_{km} \equiv \langle \lambda_k \rho_m \rangle, \quad t_{mn} \equiv \langle \rho_m \rho_n \rangle; \tag{10}$$

and the second-layer weights of the students. In other words, asymptotically

$$\lim_{D\to\infty} \epsilon^\dagger(\mathbf{W}, \mathbf{W}^\dagger, \mathbf{h}^\dagger, \mathbf{v}^\dagger) = \epsilon^\dagger(\mathbf{Q}, \mathbf{R}, \mathbf{T}, \mathbf{h}^\dagger, \mathbf{v}^\dagger). \tag{11}$$

where $\mathbf{Q} = (q_{kl})$, etc. Note there is an equivalent formulation for ‡ with the η local field and relevant second-layer weights. These overlap matrices, or "order parameters" in statistical physics jargon, have a clear physical interpretation, which can be seen when evaluating the averages explicitly. The so-called teacher-student overlap, r_{km} for example:

$$r_{km} \equiv \langle \lambda_k \rho_m \rangle = \frac{\mathbf{w}_k \mathbf{w}_m^\dagger}{D}, \tag{12}$$

quantifies the overlap or similarity between the weights of the k^{th} hidden unit of the student and the m^{th} hidden unit of the teacher. Similarly, q_{kl} gives the self-overlap of the kth and lth student nodes, and t_{mn} gives the (static) self-overlap of teacher nodes.

Task similarity The teacher-student setup gives us precise control over the task similarity via the overlap between the first-layer weights of different teachers,

$$v_{mp} \equiv \langle \rho_m \eta_p \rangle = \frac{1}{D} \mathbf{w}_m^\dagger \mathbf{w}_p^\ddagger, \tag{13}$$

which we can tune to observe its effects on the dynamics of continual learning.

2.1. Results

2.1.1. An asymptotically exact theory for the dynamics of continual learning

The test error Eq. 2 can be written in terms of the order parameters Eq. 10, so to compute the test error at all times we need to describe the evolution of $\mathbf{Q}$ etc. during training of the student using SGD Eq. 5. Such equations were derived in the vanilla teacher-student setup by Saad & Solla (1995a); Riegler & Biehl (1995), and here we extend this approach to our continual learning model. We illustrate their derivation for the teacher-student overlap r_{km} Eq. 12. Taking the inner product of Eq. 5a (in the case of $* = \dagger$) with $\mathbf{w}_n^\dagger$ and substituting the SGD update Eq. 5a yields

$$\mathrm{d}r_{km}^\mu \equiv r_{km}^{\mu+1} - r_{km}^\mu = \frac{\mathbf{w}_k^{\mu+1} \mathbf{w}_m^\dagger}{D} - \frac{\mathbf{w}_k^\mu \mathbf{w}_m^\dagger}{D} \tag{14}$$

$$= -\frac{\alpha_{\mathbf{W}}}{D} h_k^{\dagger\mu} g'(\lambda_k^\mu) \Delta^{\dagger\mu} \rho_m^\mu \tag{15}$$

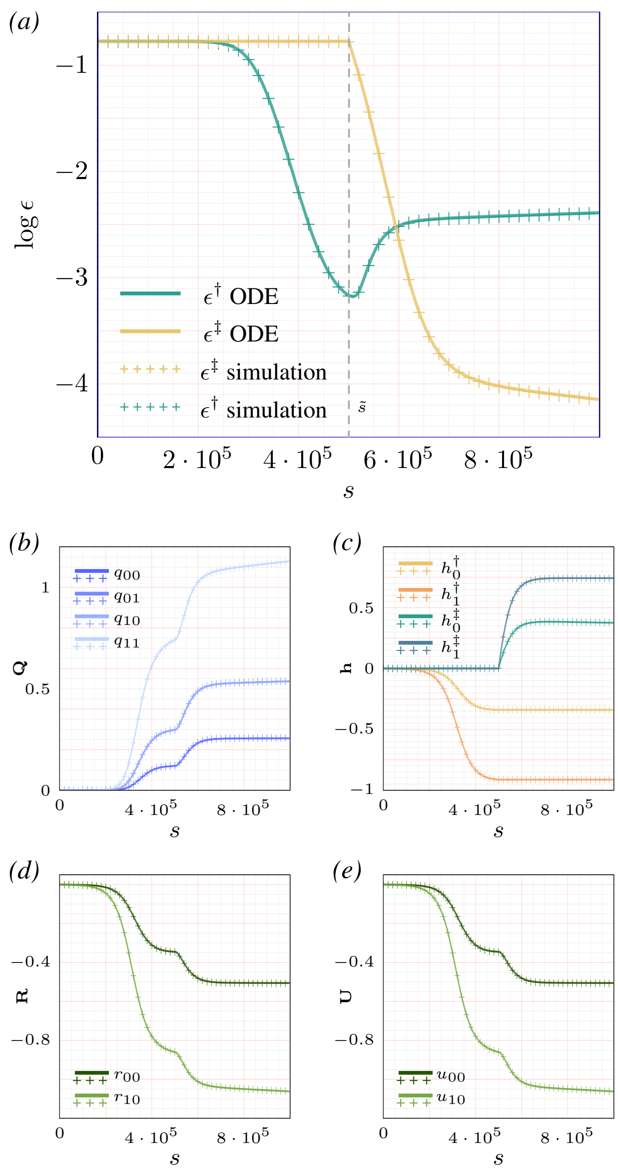

Figure 2. **Asymptotic theory matches finite-sized simulations.** Plots of progression of generalisation error (*a*) and order parameters (*b-e*) for both neural network simulations (crosses) and ODE solutions (solid lines). In this example, the teachers are fully overlapping as evidenced by the identical trajectories of $\mathbf{U}$ and $\mathbf{R}$, the student-teacher† and student-teacher‡ overlaps respectively. There is a teacher switch at step 25,000. There is close agreement between the simulation and differential equations (to within the expected $1/\sqrt{N}$ deviation).

In the thermodynamic limit $D \to \infty$, the normalised number of steps $\tau \equiv \mu/D$ becomes a continuous, time-like variable and we can write:

$$\frac{\mathrm{d}r_{kn}}{\mathrm{d}\tau} = -\alpha_{\mathbf{W}} h_k^{\dagger} \langle g'(\lambda_k) \Delta^{\dagger} \rho_n \rangle. \qquad (16)$$

The remaining averages like $\langle g'(\lambda_k)\lambda_\ell g(\rho_n)\rangle$ are simple three-dimensional integrals over the Gaussian random variables $(\lambda_k, \lambda_\ell, \rho_m)$ and can be evaluated analytically for $g(x) = \mathrm{erf}(x/\sqrt{2})$ and for linear networks with $g(x) = x$. Furthermore, these averages can be expressed only in term of the order parameters, and so the equations close. The ODEs for $\mathbf{Q}$ (Eq. D.9), $\mathbf{U}$ (Eq. D.10), as well as the student head weights, $\mathbf{h}^{\dagger}$, and $\mathbf{h}^{\ddagger}$ (Eq. D.11), are given in App. C.

In Fig. 2, we show test errors and order parameters obtained from numerically solving the ODEs (lines) and from a single simulation (crosses). We find that the agreement between theory and simulations is very good both for the test error and on the level of individual order parameters even at intermediate system size ($D = 10^4$).

2.1.2. Impact of task similarity

We integrated the ODEs in the simplest possible case to analyse the impact of task similarity. A student with $K = 2$ neurons is trained on two teachers with $M = 1$ neuron each, all having sigmoidal activation $g(x) = \mathrm{erf}(x/\sqrt{2})$[1]. A subset of the experiments was also carried out on Rectified Linear Unit (ReLU) networks (purely with network simulations) with broadly similar result; details are discussed in App. I. For $M = 1$, the task similarity $\mathbf{V}$ (Eq. 13) becomes a scalar quantity that we denote V, which is the cosine angle between the teachers' input-to-hidden weights. We parametrically generate teachers with specified similarities using the procedure described in App. F. The teacher head weights are ± 1 and the input-hidden weights are normalised. For odd activation functions like the scaled error function, the sign of the input-to-hidden weights can be compensated for by the readout weights so it is sufficient to show results for $V \in [0, 1]$. Note that the student has enough capacity to learn both teachers. Fig. 3a shows the generalisation error of the student on the first task, which decays exponentially after an initial period of stationary error. After the switch at $\tilde{s} = 1 \times 10^6$, the learning curves separate depending on the task similarity.

We plot the **forgetting** Eq. 3 at different times after the switch vs. V in panel c. For teachers with orthogonal first-layer weights ($V = 0$), performance on the first task degrades after an initial period of no forgetting. For identical first-layer weights ($V = 1$), the initial rate of forgetting is large, but the student recovers and the error on the first task plateaus at a relatively low value. In both cases, forgetting is small compared to teachers with intermediate correlations. Our model thus reproduces the empirical findings on deep networks for image classification from Ramasesh et al. (2020). We hypothesise that while it is intuitive that similar teachers lead to small forgetting, orthogonal teachers can be

[1] Code for all experiments and ODE simulations can be found at https://github.com/seblee97/student_teacher_catastrophic

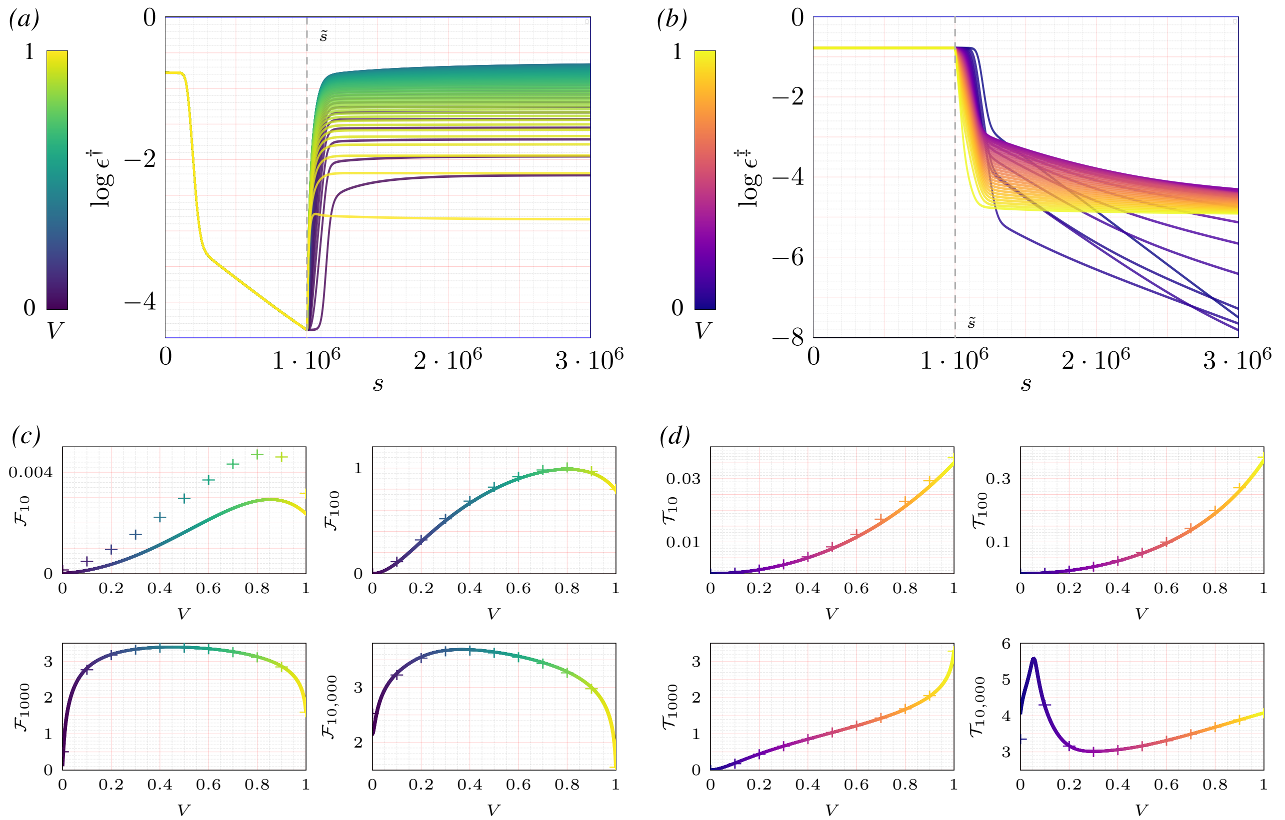

Figure 3. **The impact of task similarity on continual learning in the large-input limit**. *a (b)*: Generalisation error (2) with respect to first (second) teacher (2) during training on two teachers of overlap V computed from the ODEs of Sec. 2. In the first phase of training the generalisation error trivially follows the same trajectory for all teacher-teacher overlaps since the student only knows about one teacher. Post-switch, the generalisation errors follows different paths depending on the relationship between teachers. *c (d)*: Aggregate forgetting (transfer), $\mathcal{F}_t$ ($\mathcal{T}_t$), vs. teacher-teacher overlap, V, at various intervals after task-switch. $V = 0$ corresponds to orthogonal teacher weight vectors, $V = 1$ corresponds to identical teacher weight vectors. Here we plot with crosses the results achieved with network simulations on top of the lines representing the ODE solutions. Forgetting is strongest for teachers that are intermediately correlated, while the student is relatively robust to forgetting for aligned or orthogonal teachers. Transfer is initially monotonically better for higher overlaps; in the long time limit higher overlap appears to lead to long saddle points that are avoided in lower overlap regimes. Here and throughout, graphs with cold colour tones (greens to blues) refer to the first teacher (†), while graphs with warm colour tones (yellows to reds) refer to the second teacher (‡). The deviation of theory and simulation in the top left of (c) is a finite-size effect; the deviation is smaller than $1/\sqrt{N}$. *Parameters:* $N = 10000, M = 1, K = 2$.

more easily separated by the student by specialising to the respective teacher units. This separation is made harder by correlations between the weights of the teachers, akin to the problem of source separation in signal processing.

To analyse **transfer**, we look at the generalisation on the second task ‡ after the switch Fig. 3b. Just after the switch, higher overlap allows faster transfer. All students then reach a second plateau. Only students trained on tasks that are close to orthogonal break away form this second plateau and achieve an exponentially decaying generalisation error, whereas the other students remain stuck. This is a remarkable result, since the same student starting from random initial weights would converge to the second teacher without problem. Indeed, for orthogonal teachers, a student trained to convergence on the first task will have the equivalent of random initial weights for the second task (up to the scaling of the networks), explaining its better performance. Students trained on correlated tasks also converge to the teachers of the first, but this correlated initialisation leads to a loss of performance on the second task. We thus find that task similarity aids short-term transfer but harms long-term transfer in this setting.

3. Disentangling Feature and Read-out Similarity

In the previous sections, the task similarity is measured by the teacher-teacher overlap V, which is a metric over the input to hidden weights of the teachers. There is however another notion of task similarity that is relevant for two-layer networks: the read-out similarity, which is a metric over the hidden to output weights. A diagram showing the distinction between these similarities for a toy image task is shown in Fig. 4. Most previous studies (Goodfellow et al., 2013; Ramasesh et al., 2020) have not directly studied this distinction, although Ramasesh et al. (2020) provided evidence that the layers of a deep network that are closer to the input are more responsible to forgetting, pointing to the fact that different layers in a network might have different impact on forgetting. The teacher-student framework allows us to disentangle these different aspects of task similarity in detail.

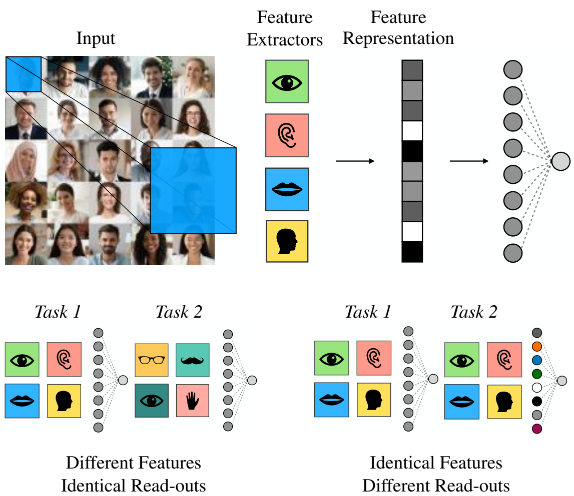

Figure 4. **Distinction between feature similarity and read-out similarity**. In a toy image task, a model consists of a set of feature extractors, which produce a feature representation vector from the input. This representation is fed into a subsequent *readout* function (top). This distinction between feature extractors and readouts leads to two related notions of task similarity: feature similarity (bottom left), which is captured by $\mathbf{V}$ in our framework Eq. 13, and read-out similarity, which we describe using $\tilde{V}$ Eq. 18. Some tasks may require the same features with different read-outs while others require different features but similar read-outs. For concrete examples of these similariity notiions, consider a continual learning task setting involving binary classification on MNIST images where the first task requires distinguishing between odd digits and even digits and the second requires distinguishing between digits less than or equal to 4 and digits greater than 4. Both tasks could be achieved with the same features but different readout functions. On the other hand two tasks where the first consists of counting blue squares and the second counting red circles will require different features but could use the same readout function.

3.1. The Mean-Field Limit of Neural Networks

In the large input-limit networks we were considering previously, the hidden dimensions were small and there was no meaningful way of defining a similarity over the hidden to output weights. For this reason, in this section we consider networks in the *mean-field limit* of large hidden dimension,

$$\phi(\mathbf{x};\mathbf{W},\mathbf{v}) = \frac{1}{M}\sum_{m}^{M} \mathbf{v}_m g(\mathbf{W}_m \mathbf{x}), \tag{17}$$

where we let the number of neurons $M \to \infty$ while the input dimension D remains finite. Note the scaling is different here from the definition in Eq. 1. In analogy to Eq. 13, let us define the teacher-teacher read-out similarity as:

$$\tilde{V} = \mathbf{h}^{\dagger} \cdot \mathbf{h}^{\ddagger}. \tag{18}$$

We can thus control both the feature and readout similarities of the teachers, and measure forgetting and transfer of the student in the $(V, \tilde{V})$ plane. However, the student dynamics cannot be described by the simple set of ODEs from above; instead, the dynamics of the student are captured by the time-evolution of the density $\rho(\boldsymbol{\theta})$ of the full set of parameters $\boldsymbol{\theta}$ of the network (Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2018; Chizat & Bach, 2018; Sirignano & Spiliopoulos, 2019). This density obeys a partial differential equation,

$$\partial_t \rho_t(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \cdot \left(\rho_t(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \Psi\left(\boldsymbol{\theta}; \rho_t\right)\right) \tag{19}$$

where $\Psi(\boldsymbol{\theta})$ can be thought of as a potential for the dynamics. This PDE is hard to analyse, so for the remainder of the paper, we resort to numerical experiments. Since the output dimension of the regression tasks is 1, we can construct teacher readout weights for a given $\tilde{V}$ with the same procedure as was used for the feature weights in previous sections. For the feature weights in the mean-field regime, we first draw a weight matrix for the first teacher element-wise i.i.d from the normal distribution and orthonormalise it: $\mathbf{W}^{\dagger} = (\mathbf{w}^{\dagger}) \in \mathbb{R}^{D\times M}$. For the second teacher, the feature weights are obtained from

$$\mathbf{W}^{\ddagger} = \alpha \mathbf{W}^{\dagger} + (1-\alpha^2)\mathbf{Z}, \tag{20}$$

where $\mathbf{Z}$ is another $D \times M$ matrix with i.i.d. Gaussian elements, and $0 < \alpha < 1$ is an interpolation parameter. It can be easily verified that α also can be interpreted as an overlap between the two feature weight matrices, where 0 corresponds to orthogonal features and 1 corresponds to identical features. To make this link clear, we abuse notation slightly and refer to α as V in analogy with previous sections.

3.2. Results

Our results are presented in Fig. 5, where we show (i) the initial rate of forgetting/transfer, (ii) the maximum amount of forgetting/transfer and (iii) the long-time values of forgetting/transfer (the value measured at the end of our training). Details of the procedure used for computing these measures are given in App. L.

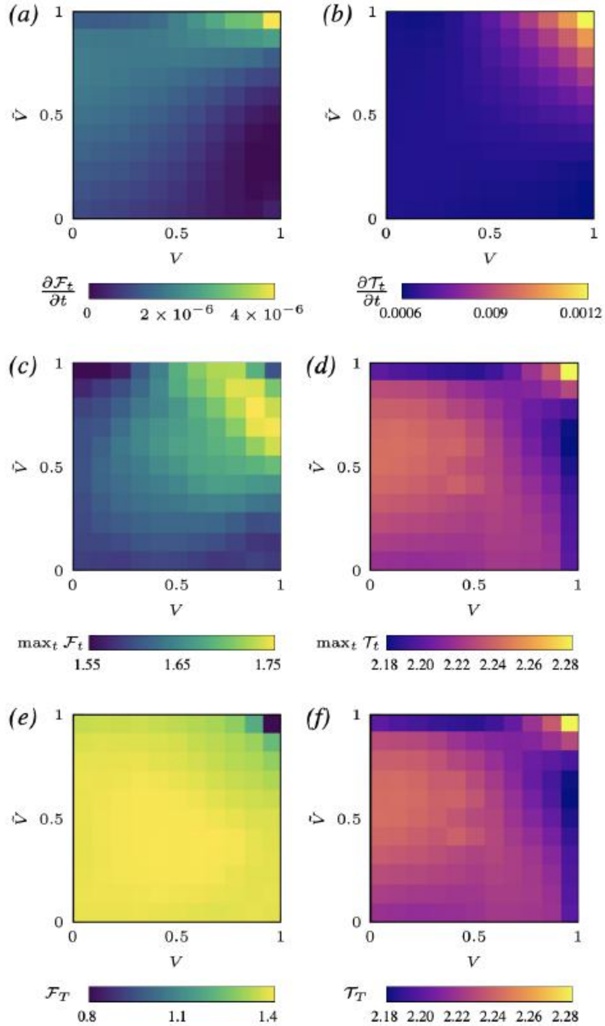

Figure 5. **The interplay of transfer and forgetting as a function of feature and readout similarity.** *(a):* Initial rate of forgetting, $\partial\mathcal{F}_t/\partial t$; *(b):* Initial rate of transfer, $\partial\mathcal{T}_t/\partial t$; *(c):* Aggregate (long-time) forgetting, $\mathcal{F}_T$; *(d):* Aggregate (long-time) transfer, $\mathcal{T}_T$; *(e):* Max forgetting reached, $\max_t \mathcal{F}_t$; *(f):* Max transfer reached, $\max_t \mathcal{T}_t$. All quantities are plotted on a 2D grid of readout similarity $\tilde{V}$ vs. feature similarity V in two-layer networks in the mean-field limit. *Parameters:* $N = 15, M = 1000, K = 250$.

As a check on our setup, we observe that our results are consistent with the preceding ODE limit simulations, as can be seen from the non-monotonic amount of forgetting for the maximum forgetting metric as a function of input similarity V for full readout similarity ($\tilde{V} = 1$, Fig. 5c top row); and the same positive relationship between feature similarity and transfer observed in the early phase of transfer in the ODE limit (Fig. 5d top row). We now turn to describing behaviour in the full space of feature and readout similarities.

3.2.1. Initial Forgetting & Transfer Rate

First, we examine the rate of forgetting and transfer at the moment the tasks are switched (Fig. 5a-b). The transfer rate is approximately constant along each diagonal, such that it is an approximate function of the summed feature and readout similarity $\tilde{V} + V$, indicating that both similarities are roughly interchangeable. By contrast, the forgetting rate shows a differential effect of each similarity type, with high readout similarity causing faster forgetting.

3.2.2. Max & Long-Time Forgetting & Transfer

The results in Fig. 5c-f demonstrate an intricate relationship between each type of task similarity and forgetting/transfer dynamics. We make several observations. First, the maximum and long-time metrics differ substantially for forgetting. For instance, the best scenario for limiting maximum forgetting is orthogonal features and fully aligned readouts ($V = 0, \tilde{V} = 1$), whereas for long-time forgetting it is complete task overlap ($\tilde{V} = 1, V = 1$). Intuitively, learning the same task twice might cause transient forgetting, but eventually will converge to the correct features for both tasks. Maximum forgetting is worst in a narrow band of high summed similarity, whereas long-time forgetting is worst at more moderate levels of summed similarity. Intuitively, very similar but subtly different tasks can produce large transient errors which are ultimately corrected at long times. Finally, for a fixed summed similarity $\tilde{V} + V$, forgetting is worst when both similarities are approximately equal. This finding generalises the observation that intermediate task similarity is most harmful to forgetting. For transfer, by contrast, the maximum and long-time metrics are highly correlated and often coincide (the point of maximum transfer is the long-time cutoff in our experiment). Outside of tasks that overlap completely, there is a slight trend for better transfer at moderate readout similarity and low feature similarity.

For a constant feature or readout similarity (that is, isolating any column or row with fixed V or $\tilde{V}$ respectively), we typically observe a non-monotonic relationship between similarity and forgetting that peaks at some intermediate level of similarity. Hence the finding that intermediate amounts of similarity cause greatest forgetting, observed in the ODE limit, holds true at most fixed similarities (see App. M for details).

Finally, we note that transfer depends on readout similarity even for teachers with identical features. Readout similarity has a non-monotonic effect, as can be seen in the column corresponding to full feature similarity ($V = 1$). This finding occurs despite the fact that the student network uses a distinct readout head for each task. We surmise that the readout weights bias the solution found by the student for the feature weights during training on the first task. This bias can have the effect of favouring subsequent learning with respect to a second teacher with readout weights that are more similar to those of the first. We show further empirical evidence for this phenomenon in App. N.

4. Conclusion

Overall, our results depict a complex relationship between task similarity, forgetting, and transfer dynamics. The degree of readout and feature similarity, as well as the timing and form of measurements, all matter in determining the outcome even qualitatively. By characterising the continual learning behaviour of simple gradient descent, we hope our experiments and theoretical framework will serve as a useful foundation for future investigations into the effect of proposed solutions for continual learning in the teacher-student setup, ultimately leading to improved algorithms.

Acknowledgements

We acknowledge support from a Sir Henry Dale Fellowship to A.S. from the Wellcome Trust and Royal Society (grant number 216386/Z/19/Z). A.S. is a CIFAR Azrieli Global Scholar in the Learning in Machines & Brains programme.

References

Advani, M., Saxe, A., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428 – 446, 2020.

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019.

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. *arXiv preprint arXiv:2105.07385*, 2021.

Aubin, B., Maillard, A., Barbier, J., Krzakala, F., Macris, N., and Zdeborová, L. The committee machine: Computational to statistical gaps in learning a two-layers neural network. In *Advances in Neural Information Processing Systems 31*, pp. 3227–3238, 2018.

Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S., Sohl-Dickstein, J., and Ganguli, S. Statistical Mechanics of Deep Learning. *Annual Review of Condensed Matter Physics*, 11(1):501–528, 2020.

Baity-Jesi, M., Sagun, L., Geiger, M., Spigler, S., Arous, G., Cammarota, C., LeCun, Y., Wyart, M., and Biroli, G. Comparing Dynamics: Deep Neural Networks versus Glassy Systems. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Bennani, M. A. and Sugiyama, M. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020.

Biehl, M. and Schwarze, H. Learning drifting concepts with neural networks. *Journal of Physics A: Mathematical and General*, 26(11):2651, 1993.

Biehl, M. and Schwarze, H. Learning by on-line gradient descent. *J. Phys. A. Math. Gen.*, 28(3):643–656, 1995.

Biehl, M., Riegler, P., and Wöhler, C. Transient dynamics of on-line learning in two-layered neural networks. *Journal of Physics A: Mathematical and General*, 29(16), 1996.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems 31*, pp. 3040–3050, 2018.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems 33*, pp. forthcoming, 2019.

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

Dhifallah, O. and Lu, Y. M. Phase transitions in transfer learning for high-dimensional perceptrons. *arXiv preprint arXiv:2101.01918*, 2021.

Doan, T., Bennani, M., Mazoure, B., Rabusseau, G., and Alquier, P. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. *arXiv preprint arXiv:2010.04003*, 2020.

Du, S., Lee, J., Tian, Y., Singh, A., and Poczos, B. Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1339–1348, 2018.

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

Farquhar, S. and Gal, Y. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018.

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.

Gabrié, M. Mean-field inference methods for neural networks. *Journal of Physics A: Mathematical and Theoretical*, 53(22):223002, 2020.

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.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems 32*, pp. 9111–9121, 2019.

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.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

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.

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.

Lampinen, A. K. and Ganguli, S. An analytic theory of generalization dynamics and transfer learning in deep linear networks. *arXiv preprint arXiv:1809.10374*, 2018.

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

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.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Mirzadeh, S. I., Farajtabar, M., Pascanu, R., and Ghasemzadeh, H. Understanding the role of training regimes in continual learning. *arXiv preprint arXiv:2006.06958*, 2020.

Ndirango, A. and Lee, T. Generalization in multitask deep neural classifiers: a statistical physics approach. In *Advances in Neural Information Processing Systems*, pp. 15862–15871, 2019.

Neyshabur, B., Sedghi, H., and Zhang, C. What is being transferred in transfer learning? In *Advances in neural information processing systems*, volume 33, 2020.

Nguyen, C. V., Achille, A., Lam, M., Hassner, T., Mahadevan, V., and Soatto, S. Toward understanding catastrophic forgetting in continual learning. *arXiv preprint arXiv:1908.01091*, 2019.

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

Ramasesh, V. V., Dyer, E., and Raghu, M. Anatomy of catastrophic forgetting: Hidden representations and task semantics. *arXiv preprint arXiv:2007.07400*, 2020.

Riegler, P. and Biehl, M. On-line backpropagation in two-layered neural networks. *Journal of Physics A: Mathematical and General*, 28(20):L507, 1995.

Rotskoff, G. and Vanden-Eijnden, E. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in neural information processing systems*, pp. 7146–7155, 2018.

Ruder, S. and Plank, B. Learning to select data for transfer learning with bayesian optimization. *arXiv preprint arXiv:1707.05246*, 2017.

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, 1995a.

Saad, D. and Solla, S. A. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225, 1995b.

Saxe, A., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B., and Cox, D. On the information bottleneck theory of deep learning. In *ICLR*, 2018.

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

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.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 2019.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Soltanolkotabi, M., Javanmard, A., and Lee, J. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Tian, Y. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pp. 3404–3413, 2017.

Toneva, M., Sordoni, A., Combes, R. T. d., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Watkin, T., Rau, A., and Biehl, M. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2): 499–556, 1993.

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.

Zdeborová, L. Understanding deep learning is also a job for physicists. *Nature Physics*, 2020.

Zdeborová, L. and Krzakala, F. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

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.

Zhong, K., Song, Z., Jain, P., Bartlett, P., and Dhillon, I. Recovery guarantees for one-hidden-layer neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4140–4149, 2017.

Zimmer, M., Viappiani, P., and Weng, P. Teacher-student framework: a reinforcement learning approach. In *AAMAS Workshop Autonomous Robots and Multirobot Systems*, 2014.