A THEORY FOR KNOWLEDGE TRANSFER IN CONTINUAL LEARNING

Diana Benavides-Prado
School of Computer Science
The University of Auckland
New Zealand
d.benavides-prado@auckland.ac.nz

Patricia Riddle
School of Computer Science
The University of Auckland
New Zealand
p.riddle@auckland.ac.nz

ABSTRACT

Continual learning of a stream of tasks is an active area in deep neural networks. The main challenge investigated has been the phenomenon of catastrophic forgetting or interference of newly acquired knowledge with knowledge from previous tasks. Recent work has investigated forward knowledge transfer to new tasks. Backward transfer for improving knowledge gained during previous tasks has received much less attention. There is in general limited understanding of how knowledge transfer could aid tasks learned continually. We present a theory for knowledge transfer in continual supervised learning, which considers both forward and backward transfer. We aim at understanding their impact for increasingly knowledgeable learners. We derive error bounds for each of these transfer mechanisms. These bounds are agnostic to specific implementations (*e.g.* deep neural networks). We demonstrate that, for a continual learner that observes related tasks, both forward and backward transfer can contribute to an increasing performance as more tasks are observed.

1 INTRODUCTION

Learning a continual stream of tasks has been a long-standing challenge in machine learning (Ring, 1997; Chen & Liu, 2018). Continual learning with deep neural networks has been an active area of research over the past few years (Delange et al., 2021), and it has multiple applications in a range of problem domains (Lesort et al., 2020; Lee & Lee, 2020; Maschler et al., 2021). Catastrophic forgetting of existing knowledge for tasks learned sequentially has been the main challenge (Delange et al., 2021). A variety of methods for this problem in supervised continual learning have been proposed, including approaches for replaying examples (Lopez-Paz & Ranzato, 2017), regularisation-based methods (Kirkpatrick et al., 2017) and network expansion methods (Ostapenko et al., 2019).

Knowledge transfer has recently been explored as an alternative for improving the performance of continual learning systems. Transferring knowledge in the forward direction has demonstrated some gains (Ke et al., 2021). Backward transfer on the other hand has been paid much less attention in continual learning with deep neural networks (Riemer et al., 2018; Ke et al., 2020; Vogelstein et al., 2020; New et al., 2022). However, backward transfer has succeeded in other lifelong learning studies that use techniques such as Support Vector Machines (SVMs) (Benavides-Prado et al., 2020), and continues to be a desired property of continual learning systems (Rish, 2022).

We develop a theory for knowledge transfer in continual learning. We first derive error bounds for individual tasks, when these are subject to forward transfer when learned for the first time, or to backward transfer from future tasks when these are learned. We then consider the order of arrival of tasks, since this influences the the amount of transfer that task is subject to. Based on the bounds derived for individual tasks, we calculate error bounds for a continual learner that learns related tasks sequentially using forward and backward transfer.

Our framework relies on three core assumptions. First, the continual learner is embedded into an *environment* of related tasks. This allows us to treat the problem of learning a sequence of tasks as the problem of learning a bias for the whole environment incrementally. Learning this bias is helpful since the continual learning will perform better at any task in that environment. Our second assumption is that relatedness between these tasks relies on the *similarity* between their example generating distributions. This assumption allows us to use a set of *transformation functions* as a tool for constraining the hypothesis family for learning a particular task, based on its similarity to other tasks in the environment (from which forward or backward transfer are to be performed). This tool has been used in other studies in multitask learning (Ben-David & Borbely, 2008). Our final assumption is that each task has a sufficient number of examples from which to learn. This assumption distinguishes our framework from approaches in zero-shot

or few-shot learning. However, in Section 4 we demonstrate that the number of examples required to learn decreases with the number of tasks.

Our proposed framework is generic and does not rely on any practical assumptions about the implementation of the continual learner (*e.g.* in terms of the learning technique, the implementation architecture, the mechanisms used for transfer or the continual learning scenario - task-incremental, domain-incremental or class-incremental (Van de Ven & Tolias, 2019)). We aim to provide a rigorous theoretical analysis to show the potential of knowledge transfer while learning sequentially, and to encourage more research in this direction.

This paper is organised as follows. Section 2 describes previous research in knowledge transfer for continual learning. Section 3 provides preliminaries and notation. Section 4 describes error bounds derived for tasks learned continually using forward knowledge transfer. Section 5 describes error bounds for tasks learned continually using backward transfer. Section 6 describes error bounds for a continual learner that uses both forward and backward transfer. Finally, Section 8 provides some discussion and final remarks.

2 Previous Research

Catastrophic forgetting or interference of new tasks with previously acquired knowledge has been studied extensively in supervised continual learning with deep neural networks (Delange et al., 2021). Several methods to avoid catastrophic forgetting have been proposed, ranging from example replay (Lopez-Paz & Ranzato, 2017; van de Ven et al., 2020) to regularisation-based (Kirkpatrick et al., 2017; Zeng et al., 2019) to dynamic networks (Yoon et al., 2017; Hung et al., 2019). Beyond catastrophic forgetting, the classic aim of continual learning systems has been to achieve increasingly knowledgeable systems (Ring, 1997; Chen & Liu, 2018). Knowledge transfer has been proposed as a mechanism to achieve this (Ke et al., 2020; Rostami et al., 2020; Benavides-Prado, 2020). Forward transfer with continual deep neural networks has been studied recently (Ke et al., 2021). Backward transfer, in contrast, has received much less attention (Riemer et al., 2018; Ke et al., 2020; Vogelstein et al., 2020), although it was explored with alternative techniques such as SVM (Benavides-Prado et al., 2020).

Ben-David and Borbely (2008) and Baxter (2000) studied the effects of learning multiple related tasks jointly with multitask learning. Baxter (2000) derived the expected average error for a group of tasks learned jointly. Ben-David and Borbely (2008) derived similar bounds for a single task learned under the same framework. More recently, Benavides-Prado, Koh and Riddle (2020) derived error bounds of knowledge transfer across SVM models in supervised continual learning. This research showed that given a set of related tasks, backward transfer with SVM can be used to achieve systems that improve their performance with each incoming task (Benavides-Prado et al., 2020). Furthermore, forward transfer can also be used to aid learning of new tasks. Although novel, these bounds were specific to the implementation using SVM. Here we extend this work by deriving error bounds that are agnostic to the implementation, for both forward transfer and backward transfer. We also derive error bounds for a continual learner that uses knowledge transfer whilst learning related tasks sequentially.

Other theoretical frameworks in transfer learning have studied how the degree of relatedness among tasks helps transfer (Lampinen & Ganguli, 2018), and how transfer helps curriculum learning (Weinshall et al., 2018). Theoretical studies in continual learning have studied the effects of task similarity in catastrophic forgetting (Lee et al., 2021), and discovered that optimal continual learning is NP-hard and requires perfect memory (Knoblauch et al., 2020). However, to the best of our knowledge there is no prior study that evaluates the effects of forward and backward knowledge transfer in learning a continual stream of supervised tasks.

3 Preliminaries and Definitions

Supervised continual learning is about learning a stream of tasks $\mathbf{T} = \{T_1, \ldots, T_n\}$. A given task T in the sequence has an underlying probability distribution $P(X, Y)$ (or simply P, which we use later indistinctly). For that task, the aim is to learn a function $f : X \rightarrow Y$, that maps the input space X to the output space Y. Learning works by exploring a hypothesis space H on that task, and finding the hypothesis $h \in H$ such that:

$$Er^P(h) = min_{h \in H} \mathcal{L}(h(x), y) \tag{1}$$

where $\mathcal{L}$ is a loss function. Naturally, estimating the error of h over the actual distribution P is difficult since P can not be observed directly. Instead, a sample S of m examples extracted repeatedly from P is used such that:

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \tag{2}$$

And the empirical error of $h \in H$ over S is such that:

$$\hat{Er}^{S}(h) = min_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(h(x_i), y_i) \tag{3}$$

To find the best h that satisfies Eq. 3, the learner aims to find the hypothesis that best fits this sample better, such that:

$$h^* = \inf_{h \in H} \mathcal{L}(h(x), y) \tag{4}$$

Knowledge transfer for continual learning aims to share knowledge across tasks observed sequentially. In our framework, we distinguish two types of transfer: 1) forward transfer, which aims to learn new or *target* tasks better or faster by transferring knowledge gained during tasks learned earlier, and 2) backward transfer, which aims to improve future performance over previous or *source* tasks by using knowledge collected while learning new tasks. We assume that tasks observed by the continual learner are related. Therefore, these tasks are assumed to belong to the same *environment*, and the continual learner can become better at learning in this environment as more tasks are observed.

Formally, we define the environment of the continual learner as follows:

Definition 1. *An environment $(\mathcal{P}, \mathcal{Q})$ of related tasks, corresponds to the set of all probability distributions on $\mathcal{X} \times \mathcal{Y}$, denoted $\mathcal{P}$, and a distribution on $\mathcal{P}$, denoted $\mathcal{Q}$. Instead of exploring a single hypothesis space, the continual learner has access to a family of hypothesis spaces $\mathbb{H} = \{H_1, H_2, \ldots, H_n\}$, one for each task. In practice, the learner has access to multiple samples to learn from, one sample for each task, such that $\mathbf{S} = \{S_1, \ldots, S_n\}$ are drawn at random from underlying probability distributions $\mathbf{P} = \{P_1, \ldots, P_n\}$.*

Access to a family of hypothesis spaces rather than a single hypothesis space, as in single-task learning, gives the continual learner the potential to learn a good bias that can generalise well to novel tasks from the same environment. Rather than producing a hypothesis that with high probability will perform well on future examples of a particular task, by learning related tasks continually the learner will produce a hypothesis space that with high probability will perform well on future tasks within the same environment. This main result has been demonstrated in the context of multitask learning (Baxter et al., 2000), and is the main result we demonstrate in Sections 4-6 for a stream of tasks learned continually using knowledge transfer.

The notion of relatedness for tasks in the environment of the continual learner relies on the similarity of their example generating distributions (Ben-David & Borbely, 2008). Formally, given a set of transformation functions $f \in \mathcal{F}$ such that $f : \mathcal{X} \to \mathcal{X}$, tasks in the environment are $\mathcal{F}$-related if, for some fixed probability distribution over $\mathcal{X} \times Y$, if the examples in each of these tasks can be generated by applying some $f \in \mathcal{F}$ to that distribution. Therefore, we can define the equivalence relation (Raczkowski & Sadowski, 1990) $\sim_{\mathcal{F}}$ on $\mathbb{H}$, where $\mathbb{H}$ is a family of hypothesis spaces for all tasks in the environment, as follows:

Definition 2. *Let $\{P_1, \ldots, P_n\}$ be the underlying probability distributions of a set of n tasks over a domain $\mathcal{X} \times Y$. Let $\mathcal{F}$ be a set of transformations $f : \mathcal{X} \to \mathcal{X}$. Let P_1 and P_2 be related if one can be generated from the other by applying some $f \in \mathcal{F}$, such that $P_1 = f[P_2]$ (and therefore $P_2 = f^{-1}[P_1]$) or $P_2 = f[P_1]$ (and therefore $P_1 = f^{-1}[P_2]$). The samples $\{S_1, \ldots, S_n\}$ to be used during learning tasks $\{T_1, \ldots, T_n\}$ are said to be $\mathcal{F}$-related if these samples come from $\mathcal{F}$-related probability distributions.*

Let $\mathbb{H}$ be a family of hypothesis spaces over the domain $\mathcal{X} \times Y$, and $\mathbb{H}$ be closed under the action of $\mathcal{F}$. Let $\mathcal{H}$ be a family of hypothesis spaces that consist of sets of hypotheses $[h] \in \mathbb{H}$ which are equivalent up to transformations in $\mathcal{F}$. If $\mathcal{F}$ acts as a group over $\mathbb{H}$ because:

- *For every $f \in \mathcal{F}$ and every $h \in \mathbb{H}$, $h \circ f \in \mathbb{H}$, and*
- *$\mathcal{F}$ is closed under transformation composition and inverses, i.e. for every $f, g \in \mathcal{F}$, the inverse transformation, f^{-1}, and the composition, $f \circ g$ are also members of $\mathcal{F}$*

Then the equivalence relation $\sim_{\mathcal{F}}$ on $\mathbb{H}$ is defined by: $h_1 \sim_{\mathcal{F}} h_2 \iff$ there exists $f \in \mathcal{F}$ such that $h_2 = h_1 \circ f$.

Therefore this framework considers the family of hypothesis spaces $\mathcal{H} = \{[h] : [h] \in \mathbb{H}\}$, which is the family of all equivalence classes of $\mathbb{H}$ under $\sim_{\mathcal{F}}$.

The original setting of this framework is in multitask learning (Ben-David & Borbely, 2008), where the equivalence class $[h]$ for a target task is first found using samples from all tasks. This requires to first identify aspects of all tasks

that are invariant under $\mathcal{F}$. A second step restricts the learning of a particular task to selecting a hypothesis $h' \in [h]$ as the hypothesis for that task. Therefore, the target task can benefit from transfer during this second step by exploring the hypothesis space to be explored for the target task $[h]$ that contains these invariances.

In continual learning we are faced with a similar problem, but rather than learning tasks jointly these are observed sequentially. However, provided these tasks are $\mathcal{F}$-related, we can adopt a similar framework to derive error bounds of a target task that is learned with forward transfer from a set of source tasks, and of source tasks for which knowledge is updated with backward transfer from a recently learned target tas. In the following sections we develop a theory of knowledge transfer across continual tasks that use these two transfer mechanisms.

4 Forward Knowledge Transfer across Related Tasks

In this and following sections, we will use $^{(t)}$ to refer to a target task, or target probability distribution or target sample, and $^{(s)}$ to denote a source task, or source probability distribution or source sample. In forward transfer, the aim is to learn a *target* task $T^{(t)}$ helped by knowledge obtained during previous n source tasks $\{T_1^{(s)}, \ldots, T_n^{(s)}\}$, with probability distributions $P^{(t)}$ and $\{P_1^{(s)}, \ldots, P_n^{(s)}\}$ and their corresponding observed samples $S^{(t)}$ and $\{S_1^{(s)}, \ldots, S_n^{(s)}\}$. Forward transfer for a continual learner which observes $\mathcal{F}$-related tasks is defined as follows:

Definition 3. *Given classes $\mathcal{F}$ and $\mathcal{H}$, and a set of labeled samples $\{S_1^{(s)}, \ldots, S_n^{(s)}\}$ for a set of n source tasks and a labeled sample $S^{(t)}$ for a target task, in forward knowledge transfer while learning task $T^{(t)}$, the continual learner:*

1. *Has access to $[h^*] \in \mathcal{H}$, obtained as a result of minimising $\inf_{h_1,\ldots,h_n \in [h]} \sum_{i=1}^{n} \hat{\mathbb{E}} r^{S_i^{(s)}}(h_i)$ over all $[h] \in \mathcal{H}$.*
2. *Selects $h^\Diamond \in [h^*]$ that minimises $\hat{\mathbb{E}} r^{S^{(t)}}(h')$ over all $h' \in [h^*]$, and outputs $h^\Diamond$ as the hypothesis for $T^{(t)}$.*

In practice, having access to $[h^*]$ during a target task $T^{(t)}$ implies that the continual learner can access to some representation of the knowledge obtained during previous tasks (*e.g.* access to a neural network representing that knowledge). We derive error bounds for learning a target task $T^{(t)}$ helped by knowledge transfer from $\mathcal{F}$-related source tasks as follows:

Theorem 1. *Let $\{P_1^{(s)}, \ldots, P_n^{(s)}\}$ and $P^{(t)}$ be a set of $\mathcal{F}$-related probability distributions, and $\{S_1^{(s)}, \ldots, S_n^{(s)}\}$ and $S^{(t)}$ random samples representing these distributions. Let $\mathcal{F}$ and $\mathcal{H}$ be defined as in Definition 2. Let $d_{max} = sup\{VC\text{-}dim(H) : H \in \mathcal{H}\}$. Let $d_{\mathcal{H}}(n) = max_{[h] \in \mathcal{H}} VC\text{-}dim([h])$. Let $h^\Diamond$ be selected according to Definition 3. Then, for every constant ϵ_1, ϵ_2, $\delta > 0$, with $|S^{(t)}|$ and $|S_i^{(s)}|$ defined similarly to Theorem 3 in Ben-David and Borbely (2008):*

$$|S^{(t)}| \geq \frac{64}{\epsilon_1^2}\left[2d_{max} log\frac{12}{\epsilon_1} + log\frac{8}{\delta}\right] \tag{5}$$

and, for all $i \leq n$:

$$|S_i^{(s)}| \geq \frac{88}{\epsilon_2^2}\left[2d_{\mathcal{H}}(n) log\frac{22}{\epsilon_2} + \frac{1}{2} log\frac{8}{\delta}\right] \tag{6}$$

then with probability greater than $(1-\delta)$:

$$\mathbb{E} r^{P^{(t)}}(h^\Diamond) \leq \inf_{h \in \mathbb{H}} \mathbb{E} r^{P^{(t)}}(h) + 2(\epsilon_1 + \epsilon_2) \tag{7}$$

Proof. Let $h^\#$ be the best $P^{(t)}$ label predictor in $\mathbb{H}$, *i.e.* $h^\# = \arg\min_{h \in \mathbb{H}} \mathbb{E} r^{P^{(t)}}(h)$. Let $[h^*]$ be the equivalence class picked according to Definition 3. By the choice of h^*:

$$\inf_{h_1,\ldots,h_n \in [h^*]} \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbb{E}} r^{S_i^{(s)}}(h_i) \leq \inf_{h_1,\ldots,h_n \in [h^\#]} \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbb{E}} r^{S_i^{(s)}}(h_i) \tag{8}$$

By Theorem 2 in Ben-David and Borbely (2008), with probability greater than $(1-\delta/2)$:

$$\inf_{h_1,\ldots,h_n \in [h^\#]} \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbb{E}} r^{S_i^{(s)}}(h_i) \leq \mathbb{E} r^{P^{(t)}}([h^\#]) + \epsilon_1 \tag{9}$$

and:

$$\mathbb{E} r^{P^{(t)}}([h^*]) \leq \inf_{h_1,\ldots,h_n \in [h^*]} \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbb{E}} r^{S_i^{(s)}}(h_i) + \epsilon_1 \tag{10}$$

Then, combining the inequalities above, with probability greater than $(1-\delta/2)$:

$$\mathbb{E}r^{P^{(t)}}([h^*]) \leq \mathbb{E}r^{P^{(t)}}([h^{\#}]) + 2\epsilon_1 \tag{11}$$

Since $h^{\diamond} \in [h^*]$, with probability greater than $(1-\delta/2)$, $h^{\diamond}$ will have an error for $P^{(t)}$ which is within $2\epsilon_2$ of the best hypothesis there, *i.e.* $\mathbb{E}r^{P^{(t)}}([h^*])$. Therefore:

$$\mathbb{E}r^{P^{(t)}}(h^{\diamond}) \leq \mathbb{E}r^{P^{(t)}}(h^{\#}) + 2(\epsilon_1 + \epsilon_2) \tag{12}$$

□

Theorem 1 implies that, for a sufficiently large number of examples for the sources and the target tasks, forward transfer is expected to benefit learning of a target task. This result is achieved by choosing a hypothesis space for $T^{(t)}$ which is biased towards the hypothesis space learned for previous $\mathcal{F}$-related tasks from the same environment. The extent of this benefit depends on the number of examples per task (see Eq. 5 and Eq. 6). Baxter (2000) demonstrated that the number of examples required per task decreases along with an increasing number of tasks, in particular:

$$|S| = \mathcal{O}\Big(\frac{1}{n} log \mathcal{C}(\epsilon, \mathcal{H}_l^n)\Big) \tag{13}$$

where $\mathcal{C}(\epsilon, \mathcal{H}_l^n)$ is the capacity of the learner given an error ϵ and a set of n sets of loss functions $\{\mathbf{h}_l^1, \ldots, \mathbf{h}_l^n\} \in \mathcal{H}_l^n$ for the family of hypothesis spaces $\mathcal{H}$. Provided that this capacity increases sublinearly with n, the number of examples required per task will decrease with an increasing number of tasks.

The amount of transfer to a target task and therefore the extent to which the bound in Theorem 1 is satisfied depends on how many source tasks are used for transfer. Intuitively, the larger this number, the smaller the bound, since the target task will have a better bias of its environment with more related tasks having been observed, which would lead to a better hypothesis space to be selected for that task. Therefore, the later a target task is observed, the greater the opportunity for it to benefit from forward transfer. This is in accordance with previous research that demonstrated that a larger number of tasks learned continually benefits transfer (Benavides-Prado et al., 2017; 2020). Next we analyse the effect of the task order in forward transfer, and the error bounds of a target task depending on that order. Next we derive bounds for forward transfer that account for the order of the task being observed in the sequence.

Definition 4. *Given classes $\mathcal{F}$ and $\mathcal{H}$, a set of labeled samples $\{S_1^{(s)}, \ldots, S_n^{(s)}\}$ for a set of source tasks and a labeled sample $S^{(t)}$ for a target task. Let:*

- $[h_n^*] \in \mathcal{H}$ *be the result of minimising* $\inf_{h_1,\ldots,h_n \in [h_n^*]} \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbb{E}} r^{S_i^{(s)}}(h_i)$ *over all* $[h] \in \mathcal{H}$*, at time* n*.*
- $[h_{n+z}^*] \in \mathcal{H}$ *be the result of minimising* $\inf_{h_1,\ldots,h_{n+z} \in [h_{n+z}^*]} \frac{1}{n+z} \sum_{i=1}^{n+z} \hat{\mathbb{E}} r^{S_i^{(s)}}(h_i)$ *over all* $[h] \in \mathcal{H}$*, at time* $n+z$*.*
- $h_n^{\diamond} \in [h_n^*]$ *that minimises* $\hat{\mathbb{E}} r^{S^{(t)}}(h')$ *over all* $h' \in [h_n^*]$*, and outputs* $h_n^{\diamond}$ *as the hypothesis for task* $T^{(t)}$ *at time* n*.*
- $h_{n+z}^{\diamond} \in [h_{n+z}^*]$ *that minimises* $\hat{\mathbb{E}} r^{S^{(t)}}(h')$ *over all* $h' \in [h_{n+z}^*]$*, and outputs* $h_{n+z}^{\diamond}$ *as the hypothesis for task* $T^{(t)}$ *at time* $n+z$*.*

Corollary 1. *Let $\{P_1^{(s)}, \ldots, P_n^{(s)}\}$, $\{S_1^{(s)}, \ldots, S_n^{(s)}\}$ and $S^{(t)}$, $\mathcal{F}$, $\mathcal{H}$, d_{max}, $d_{\mathcal{H}}(n)$ be defined as in Theorem 1, at time n. Similarly, let $\{P_1^{(s)}, \ldots, P_{n+z}^{(s)}\}$ and $P^{(t)}$ be a set of $\mathcal{F}$-related probability distributions, $\{S_1^{(s)}, \ldots, S_{n+z}^{(s)}\}$ and $S^{(t)}$ random samples representing these distributions, at time $n+z$. Let ${h^{\diamond}}_n$ and ${h^{\diamond}}_{n+z}$ be selected according to Definition 4, at time n and $n+z$, respectively. Then, for every ϵ_1, ϵ_2, $\delta > 0$, if:*

$$|S^{(t)}| \geq \frac{64}{\epsilon_1^2}\left[2d_{max} log \frac{12}{\epsilon_1} + log \frac{8}{\delta}\right] \tag{14}$$

and, at time n, for all $i \leq n$:

$$|S_i^{(s)}| \geq \frac{88}{\epsilon_n^2}\left[2d_{\mathcal{H}}(n) log \frac{22}{\epsilon_n} + \frac{1}{2} log \frac{8}{\delta}\right] \tag{15}$$

while, at time $n+z$, for all $i \leq (n+z)$:

$$|S_i^{(s)}| \geq \frac{88}{\epsilon_{n+z}^2}\left[2d_{\mathcal{H}}(n+z) log \frac{22}{\epsilon_{n+z}} + \frac{1}{2} log \frac{8}{\delta}\right] \tag{16}$$

then with probability greater than $(1-\delta)$*:*

$$\mathbb{E}r^{P^{(t)}}_{n+z}(h^{\Diamond}_{n+z}) \leq \mathbb{E}r^{P^{(t)}}_{n}(h^{\Diamond}_{n}) + \epsilon_n + \epsilon_{n+z} \tag{17}$$

See Appendix A for the proof of this corollary. The main part of the proof in Appendix A lies in Eq. 46. Since the best hypothesis space for a larger number of tasks is better than the best hypothesis space for a smaller number of tasks in the same environment, *i.e.* the bias over the environment gets refined over time, tasks observed later in the sequence will benefit more from transfer.

Bounds in Theorem 1 and Corollary 1 depend on the difference between d_{max} and $d_{\mathcal{H}}(n)$, and $VC\text{-}dim(\mathbb{H})$, with $d_{max} = sup\{VC\text{-}dim(H) : H \in \mathcal{H}\}$ and $d_{\mathcal{H}}(n) = max_{[h]\in\mathcal{H}} VC\text{-}dim([h])$, and $d_{max} \leq d_{\mathcal{H}}(n) \leq VC\text{-}dim(\mathbb{H})$. Ben-David and Borbely (2008) showed that, for a sufficiently large number or tasks n, $d_{max} = max_{[h]\in\mathcal{H}} VC\text{-}dim([h]) = d_{\mathcal{H}}(n)$. We refer readers to Section 6 of Ben-David and Borbely (2008) for details.

5 Backward Knowledge Transfer across Related Tasks

Backward transfer works by updating a source task $T^{(s)}$ using knowledge gained during the most recent target task $T^{(t)}$. Transfer occurs from the space of a target probability distribution $P^{(t)}$, represented by a sample $S^{(t)}$, to the space of a probability distribution $P^{(s)}$ that uses a sample $S^{(s)}$ for learning that source task. In continual learning, the aim is to use $P^{(t)}$, and its corresponding sample $S^{(t)}$, to bias the update of a refined version of $P^{(s)}$ towards aspects that are invariant with $P^{(t)}$, provided these are related. Benavides-Prado, Koh and Riddle (2020), analysed the special case of two tasks, one source $T^{(s)}$ and one target $T^{(t)}$, for a specific implementation of a continual learner based on SVM. Here, we present bounds for an agnostic continual learner, as follows:

Definition 5. *Given classes* $\mathcal{F}$ *and* $\mathcal{H}$*, and a pair of labeled samples* $S^{(s)}$*,* $S^{(t)}$ *for tasks* $T^{(s)}$*,* $T^{(t)}$*, during backward transfer the continual learner:*

1. *Selects* $[h^*] \in \mathcal{H}$ *that minimises* $\inf_{h^{(s)},h^{(t)}\in[h]} \left(\hat{\mathbb{E}}r^{S^{(s)}}(h^{(s)}) + \hat{\mathbb{E}}r^{S^{(t)}}(h^{(t)})\right)$ *over all* $[h] \in \mathcal{H}$*.*
2. *Selects* $h^{\Diamond} \in [h^*]$ *that minimises* $\hat{\mathbb{E}}r^{S^{(s)}}(h')$ *over all* $h' \in [h^*]$*, and outputs* $h^{\Diamond}$ *as the hypothesis for task* $T^{(s)}$*.*

In practice, the two steps in Definition 5 could be performed sequentially or jointly. For example, selecting $[h^*]$ in the first step could be performed by jointly training an auxiliary learner with examples from both $T^{(s)}$ and $T^{(t)}$, and then transferring back this information to $T^{(s)}$ during the second step. Alternatively, both $[h^*]$ could be selected jointly while training for $T^{(s)}$ aided by $T^{(t)}$.

Based on Definition 5, in the special case of two tasks $T^{(s)}$ and $T^{(t)}$:

Theorem 2. *Let* $P^{(s)}$ *and* $P^{(t)}$ *be a set of* $\mathcal{F}$*-related probability distributions,and* $S^{(s)}$ *and* $S^{(t)}$ *random samples representing these distributions on tasks* $T^{(s)}$ *and* $T^{(t)}$ *respectively. Let* $\mathcal{F}$ *and* $\mathcal{H}$ *be defined as in Definition 2. Let* d_{max} *and* $d_{\mathcal{H}}(n)$ *be defined as in Theorem 1. Let* $h^{\Diamond}$ *be selected according to Definition 5. Then, for every* ϵ_1*,* ϵ_2*,* $\delta > 0$*, if:*

$$|S^{(s)}| \geq \frac{64}{\epsilon_1^2}\left[2d_{max}log\frac{12}{\epsilon_1} + log\frac{8}{\delta}\right] \tag{18}$$

and:

$$|S^{(t)}| \geq \frac{88}{\epsilon_2^2}\left[2d_{\mathcal{H}}(2)log\frac{22}{\epsilon_2} + \frac{1}{2}log\frac{8}{\delta}\right] \tag{19}$$

then with probability greater than $(1-\delta)$*:*

$$\mathbb{E}r^{P^{(s)}}(h^{\Diamond}) \leq \inf_{h\in\mathbb{H}} \mathbb{E}r^{P^{(s)}}(h) + 2(\epsilon_1 + \epsilon_2) \tag{20}$$

Proof. Let $h^{\#}$ be the best $P^{(s)}$ label predictor in $\mathbb{H}$, *i.e.* $h^{\#} = \arg\min_{h\in\mathbb{H}} \mathbb{E}r^{P^{(s)}}(h)$. Let $[h^*]$ be the equivalence class picked according to Definition 5. By the choice of h^*:

$$\inf_{h^{(s)},h^{(t)}\in[h^*]} \left(\hat{\mathbb{E}}r^{S^{(s)}}(h^{(s)}) + \hat{\mathbb{E}}r^{S^{(t)}}(h^{(t)})\right) \leq \inf_{h^{(s)},h^{(t)}\in[h^{\#}]} \left(\hat{\mathbb{E}}r^{S^{(s)}}(h^{(s)}) + \hat{\mathbb{E}}r^{S^{(t)}}(h^{(t)})\right) \tag{21}$$

By Theorem 2 in Ben-David and Borbely (2008), in the case of two tasks:

$$\left|\mathbb{E}r^{P^{(s)}}([\hbar]) - \inf_{h^{(s)},h^{(t)}\in[\hbar]} \frac{1}{2}(\hat{\mathbb{E}}r^{S^{(s)}}(\hbar^{(s)}) + \hat{\mathbb{E}}r^{S^{(t)}}(\hbar^{(t)}))\right| \leq \epsilon_1 \tag{22}$$

then with probability greater than $(1-\delta/2)$:

$$\inf_{\hbar^{(s)},\hbar^{(t)}\in[\hbar^{\#}]} \left(\hat{\mathbb{E}}r^{S^{(s)}}(\hbar^{(s)}) + \hat{\mathbb{E}}r^{S^{(t)}}(\hbar^{(t)})\right) \leq \mathbb{E}r^{P^{(s)}}([\hbar^{\#}]) + \epsilon_1 \tag{23}$$

and:

$$\mathbb{E}r^{P^{(s)}}([\hbar^{*}]) \leq \inf_{\hbar^{(s)},\hbar^{(t)}\in[\hbar^{*}]} (\hat{\mathbb{E}}r^{S^{(s)}}(\hbar^{(s)}) + \hat{\mathbb{E}}r^{S^{(t)}}(\hbar^{(t)})) + \epsilon_1 \tag{24}$$

Then, combining the inequalities above, with probability greater than $(1-\delta/2)$:

$$\mathbb{E}r^{P^{(s)}}([\hbar^{*}]) \leq \mathbb{E}r^{P^{(s)}}([\hbar^{\#}]) + 2\epsilon_1 \tag{25}$$

Since $\hbar^{\diamondsuit} \in [\hbar^{*}]$, with probability greater than $(1-\delta/2)$, $\hbar^{\diamondsuit}$ will have an error for $P^{(s)}$ which is within $2\epsilon_2$ of the best hypothesis there, *i.e.* $\mathbb{E}r^{P_s}([\hbar^{*}])$. Therefore:

$$\mathbb{E}r^{P^{(s)}}(\hbar^{\diamondsuit}) \leq \inf_{\hbar\in\mathbb{H}} \mathbb{E}r^{P^{(s)}}(\hbar) + 2(\epsilon_1+\epsilon_2) \tag{26}$$

□

Similar to forward transfer, these bounds depend on the difference between d_{max}, $d_{\mathcal{H}}(n)$, and $VC\text{-}dim(\mathbb{H})$. Section 4 provides details on the meaning of these parameters and their relation to each other.

The main result from Theorem 2 and its corresponding proof is that an existing source task can also benefit from knowledge acquired during a related target task. This benefit is expected to be smaller than that of transferring forward, since forward transfer benefits from multiple sources (see Eq. 10) while backward transfer benefits from a single target task (see Eq. 26). We show that doing backward transfer helps to select a better hypothesis space and therefore provides a better bound on the performance of that task (see Eq. 30). Therefore, a natural next question is whether backward transfer from a sequence of target tasks, learned one at a time, can help improve these bounds. we prove that doing backward transfer multiple times sequentially helps to decrease the error on a source task $T^{(s)}$ sequentially as well.

Definition 6. *Given classes $\mathcal{F}$ and $\mathcal{H}$, a set of labeled samples $S^{(s)}$ for a source task, and labeled samples $S_n^{(t)}$, $S_{n+1}^{(t)}$ for target tasks at times n and $n+1$. Let:*

- $[\hbar_n^{*}] \in \mathcal{H}$ *be the result of minimising* $\inf_{\hbar^{(s)},\hbar^{(t)}\in[\hbar_n^{*}]} \left(\hat{\mathbb{E}}r^{S^{(s)}}(\hbar^{(s)}) + \hat{\mathbb{E}}r^{S_n^{(t)}}(\hbar^{(t)})\right)$ *over all* $[\hbar] \in \mathcal{H}$*, at time* n*.*
- $[\hbar_{n+1}^{*}] \in \mathcal{H}$ *be the result of minimising* $\inf_{\hbar^{(s)},\hbar_n^{(t)},\hbar_{n+1}^{(t)}\in[\hbar_n^{*}]} \left(\hat{\mathbb{E}}r^{S^{(s)}}(\hbar^{(s)}) + \hat{\mathbb{E}}r^{S_n^{(t)}}(\hbar_n^{(t)}) + \hat{\mathbb{E}}r^{S_{n+1}^{(t)}}(\hbar_{n+1}^{(t)})\right)$ *over all* $[\hbar] \in \mathcal{H}$*, at time* $n+1$*.*
- $\hbar_n^{\diamondsuit} \in [\hbar_n^{*}]$ *that minimises* $\hat{\mathbb{E}}r^{S^{(s)}}(\hbar')$ *over all* $\hbar' \in [\hbar_n^{*}]$*, and outputs* $\hbar_n^{\diamondsuit}$ *as the hypothesis for task* $T^{(s)}$ *at time* n*.*
- $\hbar_{n+1}^{\diamondsuit} \in [\hbar_{n+1}^{*}]$ *that minimises* $\hat{\mathbb{E}}r^{S^{(s)}}(\hbar')$ *over all* $\hbar' \in [\hbar_{n+1}^{*}]$*, and outputs* $\hbar_{n+1}^{\diamondsuit}$ *as the hypothesis for task* $T^{(s)}$ *at time* $n+1$*.*

Corollary 2. *Let $P^{(s)}$, $P_n^{(t)}$ and $P_{n+1}^{(t)}$ be a set of $\mathcal{F}$-related probability distributions, $S^{(s)}$, $S_n^{(t)}$ and $S_{n+1}^{(t)}$ random samples representing these distributions. Let $\mathcal{F}$ and $\mathcal{H}$ be defined as in Definition 2. Let d_{max} and $d_{\mathcal{H}}(n)$ be defined as in Theorem 1. Let $\hbar_n^{\diamondsuit}$ and $\hbar_{n+1}^{\diamondsuit}$ be selected according to Definition 6. Then, for every ϵ_1, ϵ_n, ϵ_{n+1}, $\delta > 0$, if:*

$$|S^{(s)}| \geq \frac{64}{\epsilon_1^2}\left[2d_{max}log\frac{12}{\epsilon_1} + log\frac{8}{\delta}\right] \tag{27}$$

and, at time n:

$$|S_n^{(t)}| \geq \frac{8}{\epsilon_n^2}\left[2d_{\mathcal{H}}(2)log\frac{22}{\epsilon_n} + \frac{1}{2}log\frac{8}{\delta}\right] \tag{28}$$

while, at time $n+1$:

$$|S_{n+1}^{(t)}| \geq \frac{88}{\epsilon_{n+1}^2}\left[2d_{\mathcal{H}}(2)log\frac{22}{\epsilon_{n+1}} + \frac{1}{2}log\frac{8}{\delta}\right] \tag{29}$$

then with probability greater than $(1-\delta)$*:*

$$\mathbb{E}r_{n+1}^{P^{(s)}}(h_{n+1}^{\Diamond}) \leq \mathbb{E}r_{n}^{P^{(s)}}(h_{n}^{\Diamond}) + \epsilon_n + \epsilon_{n+1} \tag{30}$$

See Appendix B for the proof of this corollary. These results imply that doing backward transfer sequentially whilst learning target tasks will lead to more refined hypothesis spaces in a source task, beyond the hypothesis space learned initially (with or without forward transfer). Furthermore, this suggests that continually learning $\mathcal{F}$-related tasks while doing both forward and backward transfer can lead to a better bias over the learning environment of these tasks, *i.e.* the result demonstrated by Baxter (2000) for multitask learning, which we demonstrate in the next section.

6 Continual Learning of Related Tasks using Knowledge Transfer

Based on the bounds derived in Section 4 and Section 5, now we are ready to derive bounds of a continual learner that observes supervised related tasks sequentially while doing knowledge transfer. First, lets recall from Definition 1 that the continual learner is embedded in an environment of related tasks, $(\mathcal{P}, \mathcal{Q})$, where $\mathcal{P}$ is the set of all probability distributions on $\mathcal{X} \times Y$ and $\mathcal{Q}$ is a distribution on $\mathcal{P}$. The error of a selected hypothesis space $[h^*] \in \mathcal{H}$ for all tasks in such environment is defined as:

$$\mathbb{E}r^{\mathcal{Q}}([h^*]) = \inf_{h_1,\ldots,h_n \in [h^*]} \sum_{i=1}^{n} \mathbb{E}r^{P_i}([h_i]) \tag{31}$$

for any P drawn at random from $\mathcal{P}$ according to $\mathcal{Q}$. Let's define ϵ_f as the average ϵ when performing forward transfer to a new task, *i.e.* ϵ_f corresponds to $2(\epsilon_1 + \epsilon_2)$ in Theorem 1, averaged across all tasks. Similarly, let's define ϵ_b as the average ϵ when performing backward transfer to a new task, *i.e.* ϵ_b corresponds to $2(\epsilon_1 + \epsilon_2)$ in Theorem 2 averaged across all tasks. Although the definitions of ϵ_f and ϵ_b oversimplify the continual learner to the case of all tasks achieving roughly the same error bounds by means of transfer, this will serve to demonstrate how forward and backward transfer help to improve the bounds for the continual learner as a whole. For a task i, the error bound of applying forward and backward transfer and selecting $[h^*]$ instead of $[h^\#]$ as the hypothesis space for that task is:

$$\mathbb{E}r^{P_i}([h^*]) \leq \mathbb{E}r^{P_i}([h^\#]) + (i-1)\epsilon_f + (n-i)\epsilon_b \tag{32}$$

As demonstrated in Corollary 1 and 2, the extent to which transfer helps to improve the error bounds of a particular task i depends on the order of that task in the sequence, which in Eq. 32 impacts the total amount of transfer through $(i-1)$ for forward transfer and $(n-i)$ for backward transfer. Given Eq. 32, for a sequence of tasks n, we can define the error bounds on the environment that learns those tasks by means of transfer as follows.

Theorem 3. *Let* $\{P_1, \ldots, P_n\}$ *be a set of distributions, one for each task, drawn at random from* $\mathcal{P}$*, the set of all probability distributions on* $\mathcal{X} \times Y$*, according to* $\mathcal{Q}$*, a distribution on* $\mathcal{P}$*. Let* $\mathcal{H}$ *be the family of hypothesis spaces for* n *tasks to be learned in the environment* $\mathcal{Q}$*, according to Definition 2, with* $[h^*] \in \mathcal{H}$ *selected according to Theorem 1 and Theorem 2. Let* $\mathbb{H}$ *be the family of hypothesis spaces for* n *tasks with no transfer, and let* $[h^\#] \in \mathbb{H}$ *be selected as the hypothesis space for the* n *tasks with no transfer. If the number of tasks* n *satisfies:*

$$n \geq \max\left\{\frac{256}{\epsilon^2} log \frac{8\mathcal{C}\left(\frac{\epsilon}{32}, \mathcal{H}^*\right)}{\delta}, \frac{64}{\epsilon^2}\right\} \tag{33}$$

with $\mathcal{H}^* = \{[h^*] : h \in \mathcal{H}\}$*, i.e. the set of all hypothesis spaces in the hypothesis space family* $\mathcal{H}$ *such that each* $[h^*]$ *is defined by:*

$$[h^*](P) = \inf_{h \in \mathcal{H}} \mathbb{E}r^{P}(h) \tag{34}$$

and, for all $1 \leq i \leq n$*, the number of examples per task,* $|S_i|$ *satisfies:*

$$|S_i| \geq \max\left\{\frac{256}{n\epsilon^2} log \frac{8\mathcal{C}\left(\frac{\epsilon}{32}, \mathcal{H}_l^n\right)}{\delta}, \frac{64}{\epsilon^2}\right\} \tag{35}$$

where $\mathcal{H}_l^n = \cup_{[h^*] \in \mathcal{H}} [h^*]_l^n$ *(i.e.* $\mathcal{H}_l^n$ *is the union of all sequences of hypothesis* $[h^*] \in \mathcal{H}$*, each of size* n*, subject to loss function* l*), and with:*

$$\epsilon = \sum_{i=1}^{n}(i-1)\epsilon_f + \sum_{i=1}^{n}(n-1)\epsilon_b \tag{36}$$

then, with probability at least $(1-\delta)$, $[h^*] \in \mathcal{H}$ *will satisfy:*

$$\mathbb{E}r^{\mathcal{Q}}([\mathcal{h}^*]) \leq \mathbb{E}r^{\mathcal{Q}}([\mathcal{h}^{\#}]) + \epsilon \tag{37}$$

Proof. According to Eq. 32, for all $1 \leq i \leq n$:

$$\mathbb{E}r^{P_i}([\mathcal{h}^*]) \leq \mathbb{E}r^{P_i}([\mathcal{h}^{\#}]) + (i-1)\epsilon_f + (n-i)\epsilon_b \tag{38}$$

which leads to:

$$\inf_{\mathcal{h}_1,\ldots,\mathcal{h}_n \in [\mathcal{h}^*]} \sum_{i=1}^{n} \mathbb{E}r^{P_i}([\mathcal{h}_i^*]) \leq \inf_{\mathcal{h}_1,\ldots,\mathcal{h}_n \in [\mathcal{h}]} \sum_{i=1}^{n} \Big(\mathbb{E}r^{P_i}([\mathcal{h}_i^{\#}])) + (i-1)\epsilon_f + (n-1)\epsilon_b\Big) \tag{39}$$

with:

$$\mathbb{E}r^{\mathcal{Q}}([\mathcal{h}^*]) = \inf_{\mathcal{h}_1,\ldots,\mathcal{h}_n \in [\mathcal{h}^*]} \sum_{i=1}^{n} \mathbb{E}r^{P_i}([\mathcal{h}_i^*]) \tag{40}$$

and:

$$\mathbb{E}r^{\mathcal{Q}}([\mathcal{h}^{\#}]) = \inf_{\mathcal{h}_1,\ldots,\mathcal{h}_n \in [\mathcal{h}]} \sum_{i=1}^{n} \mathbb{E}r^{P_i}([\mathcal{h}_i^{\#}]) \tag{41}$$

then:

$$\mathbb{E}r^{\mathcal{Q}}([\mathcal{h}^*]) \leq \mathbb{E}r^{\mathcal{Q}}([\mathcal{h}^{\#}]) + \epsilon \tag{42}$$

□

Theorem 3 and its corresponding proof provide the most relevant result of our framework: learning a set of tasks continually with forward and backward transfer will lead to incrementally learning a better bias over the environment itself. Furthermore, the larger the number of tasks n, the better the bounds. This can be inferred from Eq. 39, since for a particular $\mathbb{E}r^{P_i}([\mathcal{h}_i^{\#}]))$ on the right-hand side the larger the number of tasks the larger i and n are, the larger the difference, or gap, with the left-hand side. This implies that the hypothesis space for a particular task which is selected by considering other tasks, *i.e.* via transfer, is a better hypothesis space than would be selected by learning that task in isolation. Since this occurs for all tasks in the environment, the better the bias learned over that environment will be and therefore the better future tasks will be learned, leading to an increasingly knowledgeable system.

As a final remark, note that in practice the bounds in Theorem 3 depend on the samples S_i, $1 \leq i \leq n$, drawn from the corresponding P_i probability distributions, since this is the data that can be observed while learning. For these bounds to apply, for all $1 \leq i \leq n$, the bounds between P_i and S_i must satisfy (Baxter et al., 2000):

$$\mathbb{E}r^{P_i}(h) \leq \hat{\mathbb{E}}r^{S_i}(h) + \left[\frac{32}{m}\left(dlog\frac{2\epsilon m}{d} + log\frac{2}{\delta}\right)\right]^{1/2} \tag{43}$$

with d the VC-dimension of $\mathcal{H}$ and m the number of examples. Then, with probability at least $(1-\delta)$ all $h \in \mathcal{H}$ will satisfy Eq. 43.

7 Experiments

We experiment with an example inspired by multitask learning (Zhang & Yeung, 2014). A continual learner observes a set of regression tasks to learn four functions, three of which are linear and related while one is unrelated (see Figure 1). We report six different scenarios to prove bounds presented in previous sections: 1) forward transfer from f_1 to f_2 (bound in Theorem 1), 2) backward transfer from f_2 to f_1 (bound in Theorem 2), 3) forward transfer from f_1 and f_2 to f_3 (bound in Corollary 1), 4) backward transfer from f_2 to f_1 and then from f_3 to f_1 (bound in Corollary 2), 5) forward transfer from f_1 to f_4 (bound in Theorem 1 for an unrelated task), and 6) backward transfer from f_4 to f_1 (bound in Theorem 2 for an unrelated task). Each of these six scenarios is trained and tested independently from the other scenarios. We measure the final R^2 of the three tasks learned by that continual learner. We use a neural network with 1 hidden layer of 10 units to learn in each of these scenarios. For each scenario, the task from which transfer occurs is trained only partially (*i.e.* before full convergence), while the task which is subject to transfer is trained until convergence. We generate 30 random examples for each task, for values of x between 0 and 10. We add Gaussian noise with mean 1 and standard deviation 2. We split data from each task into training (75%) and test (25%) sets. We repeat sampling, splitting, training and testing 10 times. Table 1 shows that transfer across related tasks (scenarios 1

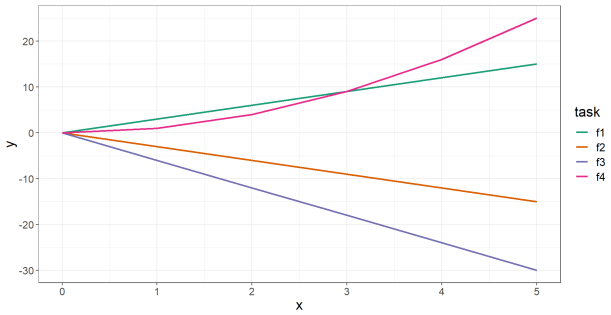

Figure 1: A set of problems or tasks. Three of these are linear functions and are related, while one is an unrelated task.

Scenario	$R^2 f_1$	$R^2 f_2$	$R^2 f_3$	$R^2 f_4$	$R^2 f_1, f_2, f_3$
Isolated learning	0.9997 ± 0.0002	0.999 ± 0.001	0.999 ± 0.001	0.9842 ± 0.0382	0.9955 ± 0.0101
Isolated learning (noise)	0.8463 ± 0.033	0.8737 ± 0.0127	0.896 ± 0.0133	0.7825 ± 0.0754	0.8496 ± 0.0034
Scenario 1 (noise)	$--$	0.8919 ± 0.0174	$--$	$--$	$--$
Scenario 2 (noise)	0.9613 ± 0.0181	$--$	$--$	$--$	$--$
Scenario 3 (noise)	$--$	$--$	0.9137 ± 0.0135	$--$	$--$
Scenario 4 (noise)	0.9636 ± 0.0172	$--$	$--$	$--$	$--$
Scenario 5 (noise)	$--$	$--$	$--$	0.4322 ± 0.0933	$--$
Scenario 6 (noise)	0.7035 ± 0.0364	$--$	$--$	$--$	$--$

Table 1: Mean R^2, and their standard deviations, of six transfer scenarios on a toy example of four functions: $y_1 = -3x + 10$, $y_2 = -3x - 5$, $y_3 = -6x - 12$ and $y_4 = x^2$ ($--$ denotes not applicable).

to 4) benefits R^2 performance. Our main finding is that, by training source tasks only partially, we are able to keep the hypothesis space large enough for the backward/forward transfer to have an effect. This also allows exploring a larger set of $\mathcal{F}$ transformation functions between hypothesis spaces, which appears to be critical for transfer. The practical implication of this is that we will need to store partially converged versions of each task's model for future transfer.

8 Discussion and Conclusion

We proposed a theory for knowledge transfer in supervised continual learning. We aim to encourage further research in knowledge transfer for achieving increasingly knowledgeable continual learning systems. Our proposed framework relies on the assumption of relatedness among tasks in a specific environment. This assumption may be applicable to a variety of domains that learn different but related tasks, *e.g.* tasks in a clinical domain, tasks in manufacturing, etc. Our error bounds are agnostic to the implementation of the continual learner. One would naturally wonder how these bounds apply to implementations with deep neural networks. Since those bounds depend on the number of examples per task, which itself depends on the number of tasks (see Eq. 13), it is possible that some data from previous tasks will be needed. Strategies such as memories per task or generative models, which have been used in several studies, could be helpful. We also believe that modular or semi-modular networks, with specialized components for each task, will potentially be necessary for effective knowledge transfer. Furthermore, having specialized modules for representing the set of transformations between hypothesis spaces for each task could be helpful. We also hypothesise that scenarios such as class-incremental learning of related classes could benefit more from transfer than scenarios such as task or domain-incremental learning, although more research is required in this direction.

Recent work suggests that the framework of the VC-dimension is not appropriate for deep neural networks. Other frameworks based on infinite-width networks (Golikov, 2020) and robustness-based networks (Bubeck et al., 2021) have been widely studied for deep neural networks that observe examples of all tasks at once. Although most work in supervised continual learning has used deep neural networks, the infinite-width and robustness-based frameworks have not been analysed with the lens of learning incrementally. Intuitively, overparameterised networks would imply a bigger challenge for continual learning, as this overparameterisation would lead to excellent performance on a particular task, making the network harder to adapt to subsequent ones. Extending studies on infinite-width or robustness-based networks to the challenges of continual learning would be an interesting avenue of research.

REFERENCES

Jonathan Baxter et al. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12(149-198):3, 2000.

Shai Ben-David and Reba Schuller Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73(3):273–287, 2008.

Diana Benavides-Prado. Beyond catastrophic forgetting in continual learning: An attempt with svm. In *ICML*. The International Conference on Machine Learning (ICML), 2020.

Diana Benavides-Prado, Yun Sing Koh, and Patricia Riddle. Accgensvm: Selectively transferring from previous hypotheses. In *Proc. Intern. Joint Conf. Artificial Intel*, pp. 1440–1446, 2017.

Diana Benavides-Prado, Yun Sing Koh, and Patricia Riddle. Towards knowledgeable supervised lifelong learning systems. *Journal of Artificial Intelligence Research*, 68:159–224, 2020.

Sébastien Bubeck, Yuanzhi Li, and Dheeraj M Nagaraj. A law of robustness for two-layers neural networks. In *Conference on Learning Theory*, pp. 804–820. PMLR, 2021.

Zhiyuan Chen and Bing Liu. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3):1–207, 2018.

Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Eugene Golikov. Towards a general theory of infinite-width limits of neural classifiers. In *International Conference on Machine Learning*, pp. 3617–3626. PMLR, 2020.

Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compacting, picking and growing for unforgetting continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Zixuan Ke, Bing Liu, and Xingchang Huang. Continual learning of a mixed sequence of similar and dissimilar tasks. *Advances in Neural Information Processing Systems*, 33:18493–18504, 2020.

Zixuan Ke, Bing Liu, Nianzu Ma, Hu Xu, and Lei Shu. Achieving forgetting prevention and knowledge transfer in continual learning. *Advances in Neural Information Processing Systems*, 34, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, pp. 201611835, 2017.

Jeremias Knoblauch, Hisham Husain, and Tom Diethe. Optimal continual learning has perfect memory and is np-hard. In *International Conference on Machine Learning*, pp. 5327–5337. PMLR, 2020.

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

Cecilia S Lee and Aaron Y Lee. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281, 2020.

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

Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, 58:52–68, 2020.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pp. 6467–6476, 2017.

Benjamin Maschler, Thi Thu Huong Pham, and Michael Weyrich. Regularization-based continual learning for anomaly detection in discrete manufacturing. *Procedia CIRP*, 104:452–457, 2021.

Alexander New, Megan Baker, Eric Nguyen, and Gautam Vallabha. Lifelong learning metrics. *arXiv preprint arXiv:2201.08278*, 2022.

Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11321–11329, 2019.

Konrad Raczkowski and Paweł Sadowski. Equivalence relations and classes of abstraction. *Formalized Mathematics*, 1(3):441–444, 1990.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.

Mark B Ring. Child: A first step towards continual learning. *Machine Learning*, 28(1):77–104, 1997.

Irina Rish. Towards general and robust ai at scale, 2022. URL https://www.youtube.com/watch?v=OL3hUZh6lTc.

Mohammad Rostami, David Isele, and Eric Eaton. Using task descriptions in lifelong machine learning for improved performance and zero-shot transfer. *Journal of Artificial Intelligence Research*, 67:673–704, 2020.

Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Gido M van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):1–14, 2020.

Joshua T Vogelstein, Jayanta Dey, Hayden S Helm, Will LeVine, Ronak D Mehta, Ali Geisa, Haoyin Xu, Gido M van de Ven, Emily Chang, Chenyu Gao, et al. Representation ensembling for synergistic lifelong learning with quasilinear complexity. *arXiv preprint arXiv:2004.12908*, 2020.

Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning*, pp. 5238–5246. PMLR, 2018.

Jaehong Yoon, Eunho Yang, et al. Lifelong Learning with Dynamically Expandable Networks. *arXiv:1708.01547*, 2017.

Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372, 2019.

Yu Zhang and Dit-Yan Yeung. A regularization approach to learning task relationships in multitask learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(3):1–31, 2014.

APPENDIX A

Proof for Corollary 1:

Proof. By Theorem 2 in Ben-David and Borbely, at time $n+z$:

$$\mathbb{E}r_{n+z}^{P^{(t)}}([h_{n+z}^*]) \leq \inf_{h_1,\ldots,h_{n+z}\in[h_{n+z}^*]} \frac{1}{n+z}\sum_{i=1}^{n+z}\hat{\mathbb{E}}r^{S_i^{(s)}}(h_i) + \epsilon_{n+z} \tag{44}$$

while, at time n:

$$\mathbb{E}r_{n}^{P^{(t)}}([h_{n}^*]) \leq \inf_{h_1,\ldots,h_{n}\in[h_{n}^*]} \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbb{E}}r^{S_i^{(s)}}(h_i) + \epsilon_{n} \tag{45}$$

And by Ben-David and Borbely (2008), and also Baxter (2000):

$$\inf_{h_1,\ldots,h_{n+z}\in[h_{n+z}^*]} \frac{1}{n+z}\sum_{i=1}^{n+z}\hat{\mathbb{E}}r^{S_i^{(s)}}(h_i) \leq \inf_{h_1,\ldots,h_{n}\in[h_{n}^*]} \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbb{E}}r^{S_i^{(s)}}(h_i) \tag{46}$$

Therefore:

$$\mathbb{E}r_{n+z}^{P^{(t)}}([h_{n+z}^*]) \leq \mathbb{E}r_{n}^{P^{(t)}}([h_{n}^*]) + \epsilon_n + \epsilon_{n+z} \tag{47}$$

With $\mathbb{E}r_{n+z}^{P^{(t)}}(h_{n+z}^\diamond)$ the best hypothesis in $\mathbb{E}r_{n+z}^{P^{(t)}}([h_{n+z}^*])$ and $\mathbb{E}r_{n}^{P^{(t)}}(h_{n}^\diamond)$ the best hypothesis in $\mathbb{E}r_{n}^{P^{(t)}}([h_{n}^*])$, the theorem is proved. □

APPENDIX B

Proof for Corollary 2:

Proof. By Theorem 2 in Ben-David and Borbely, at time $n+1$:

$$\mathbb{E}r_{n+1}^{P^{(s)}}([h_{n+1}^*]) \leq \inf_{h^{(s)}, h_n^{(t)}, h_{n+1}^{(t)} \in [h_{n+1}^*]} \left(\hat{\mathbb{E}}r^{S^{(s)}}(h^{(s)}) + \hat{\mathbb{E}}r^{S_n^{(t)}}(h_n^{(t)}) + \hat{\mathbb{E}}r^{S_{n+1}^{(t)}}(h_{n+1}^{(t)})\right) + \epsilon_{n+1} \tag{48}$$

while, at time n:

$$\mathbb{E}r_n^{P^{(s)}}([h_n^*]) \leq \inf_{h^{(s)}, h_n^{(t)} \in [h_n^*]} \left(\hat{\mathbb{E}}r^{S^{(s)}}(h^{(s)}) + \hat{\mathbb{E}}r^{S_n^{(t)}}(h_n^{(t)})\right) + \epsilon_n \tag{49}$$

And by Ben-David and Borbely (2008), and also Baxter (2000):

$$\inf_{h_1,\ldots,h_{n+1} \in [h_{n+1}^*]} \frac{1}{n+1} \sum_{i=1}^{n+1} \hat{\mathbb{E}}r^{S_i^{(s)}}(h_i) \leq \inf_{h_1,\ldots,h_n \in [h_n^*]} \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbb{E}}r^{S_i^{(s)}}(h_i) \tag{50}$$

Therefore:

$$\mathbb{E}r_{n+1}^{P^{(s)}}([h_{n+1}^*]) \leq \mathbb{E}r_n^{P^{(s)}}([h_n^*]) + \epsilon_n + \epsilon_{n+1} \tag{51}$$

With $\mathbb{E}r_{n+1}^{P^{(s)}}(h_{n+1}^\diamond)$ the best hypothesis in $\mathbb{E}r_{n+1}^{P^{(s)}}([h_{n+1}^*])$ and $\mathbb{E}r_n^{P^{(s)}}(h_n^\diamond)$ the best hypothesis in $\mathbb{E}r_n^{P^{(s)}}([h_n^*])$, the theorem is proved. □