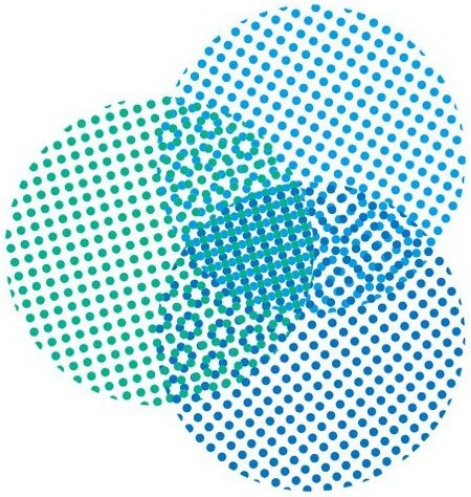

CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 071 **December 5, 2017**

Theory of Intelligence with Forgetting:
Mathematical Theorems Explaining Human Universal Forgetting using "Forgetting Neural Networks"

by

Felipe Cano-Córdoba, Sanjay Sarma, and Brian Subirana

MIT Auto-ID Laboratory

Abstract

In [46] we suggested that any memory stored in the human/animal brain is forgotten following the Ebingghaus curve – in this follow-on paper, we define a novel algebraic structure, a ***Forgetting Neural Network***, as a simple mathematical model based on assuming parameters of a neuron in a neural network are forgotten using the Ebbinghaus forgetting curve. We model neural networks in Sobolev spaces using [39] as our departure point and demonstrate four novel theorems of Forgetting Neural Networks: ***theorem of non-instantaneous forgetting, theorem of universal forgetting, curse of forgetting theorem***, and ***center of mass theorem***. We also present the possibly most efficient representation of neural networks' ***"minimal polynomial basis layer" (MPBL)*** since our basis construct can generate n polynomials of order m using only $2m+1+n$ neurons. As we briefly discuss in the conclusion, there are about 10 similarities between forgetting neural networks and human forgetting. Our research elicits more questions than it answers and may have implications for neuroscience research –including in our understanding of how babies learn (or perhaps, forget), which leads us to suggest what we call the ***baby forgetting conjecture***.

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF - 1231216.

Theory of Intelligence with Forgetting:
Mathematical Theorems Explaining Human Universal Forgetting using "Forgetting Neural Networks"

Felipe Cano-Córdoba, Sanjay Sarma, and Brian Subirana

MIT Auto-ID Laboratory

December 5, 2017

Abstract

In [46] we suggested that any memory stored in the human/animal brain is forgotten following the Ebingghaus curve – in this follow-on paper, we define a novel algebraic structure, a ***Forgetting Neural Network***, as a simple mathematical model based on assuming parameters of a neuron in a neural network are forgotten using the Ebbinghaus forgetting curve. We model neural networks in Sobolev spaces using [39] as our departure point and demon-strate four novel theorems of Forgetting Neural Networks: ***theorem of non-instantaneous forgetting, theorem of universal forgetting, curse of forgetting theorem***, and ***cen-ter of mass theorem***. We also present the possibly most efficient representation of neural networks' ***"minimal polynomial basis layer" (MPBL)*** since our basis construct can generate n polynomials of order m using only $2m + 1 + n$ neurons. As we briefly discuss in the conclusion, there are about 10 similarities between forgetting neural networks and human forgetting. Our research elicits more questions than it answers and may have implications for neuroscience research –including in our understanding of how babies learn (or perhaps, forget), which leads us to suggest what we call the ***baby forgetting conjecture***.

Reference this memo as: Cano-Córdoba, F., Sarma, S., and Subirana, B. Memo 71: Theory of Intelligence with Forgetting: Mathematical Theorems Explaining Human Universal Forgetting using "Forgetting Neural Networks". *Center for Brains, Minds and Machines* (2017). http://hdl.handle.net/1721.1/113608

Contents

1 Introduction

In this paper we aim to modify deep learning architectures so they exhibit forgetting behavior similar to that of humans [46] [1] . Notably human brains learn from few examples and forget spontaneously while deep learning networks require many examples and don't forget (unless perhaps if retrained exhibiting then what's been called in the literature "catastrophic forgetting"). If the basic mechanism of neural networks (multiply - add - activate) is also the basis for biological brains, we feel adding constraints (such as forgetting) may bring the models closer to real brains, as we have indeed found in the research reported here.

The universal law of forgetting

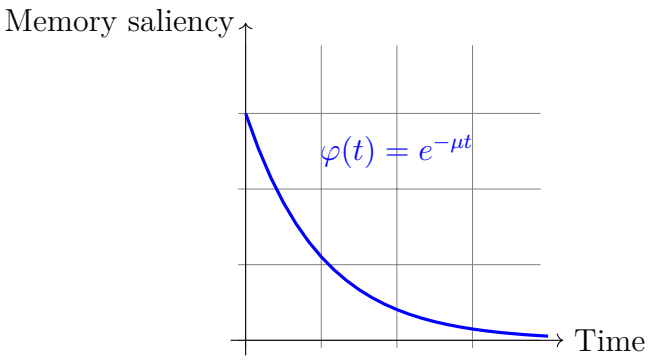

Figure 1: Saliency of "unused memory" stored in neurons follows an Ebbinghaus rule. As an additional $\ln(2)/\mu$ of time passes, memory saliency probability is cut by two. For college academics, [46] conjectures the curve is $\mu = \frac{1}{2} \cdot \ln(2)$ year^{-1}.

The extensive evidence in forgetting [46] is consistent with a common neuron-based mechanism behind all forms of memory forgetting. Given deep learning approaches have some similarities with how the natural brain may operate (multiplication of input signals, activation functions, multiple layers), we explore in this paper if, theoretically, forgetting in deep learning networks may be modeled at the neuron level too. In this paper we focus on a simple mathematical model of forgetting in Deep Neural networks based on modeling it as a function of weight loss. Our research shows this simple neural mechanism generates forgetting behavior of deep neural networks that strengthens several similarities with human memory.

Our goal is to build on existing research towards a theory of intelligence as described elsewhere ([39], [38] and [50]) by adding a very simple forgetting component to the theory. A contribution of this paper is to present in extended and pedagogical form, the key results shown in [39].

[1] A recording of a CBMM presentation of this research can be found here: http://cbmm.mit.edu/news-events/events/forgetting-college-academics-and-role-learning-engineering-building-expertise or here: https://www.youtube.com/watch?v=DdMlI6R1MJ0.

In section 2 we introduce the forgetting neurons and deep neural networks, a simple mathematical model based on assuming parameters of a neuron are forgotten using the Ebbinghaus forgetting curve, and also forgetting networks, an algebraic structure with certain basic properties inspired in human forgetting.

In section 3 we review basic mathematical concepts related to Sobolev spaces and in section 4 present two novel and basic theorems of forgetting networks: ***theorem of non-instantaneous forgetting*** and ***theorem of universal forgetting***, showing respectively that forgetting cannot happen instantaneously and that it is unavoidably fatal given sufficient time. In section 5 we give a proof of the universality theorems for shallow and deep networks and prove the ***center of mass theorem***, which states that learned knowledge is forgotten in such a way that "high" frequencies are forgotten faster, or that the knowledge tends to its center of mass and forgets the finner edges. We also present the possibly most efficient representation of polynomial basis using a neural network since our basis construct can generate n polynomials of order m using only $2m + 1 + n$ neurons. In section 6 we give theoretical bounds on network complexity to achieve a given accuracy, and present what is known as the *curse of dimensionality*. We also introduce the novel ***curse of forgetting theorem***. The first 6 sections collectively demonstrate theoretically that the simple construct used throughout the paper is a forgetting network in a Sobolev space. In section 7 we explore more complex models of forgetting, both deterministic and stochastic. In section 8 we conclude with a review of future research and show connections between the model presented and ten features of memory based on the properties of forgetting networks. Our research elicits more questions than it answers as we briefly discuss in the conclusion. The appendices contain some support proofs that have been removed from the main text for simplicity reasons.

2 Desired Constraints in Modelling Forgetting

In general, a basic description of how to perform a simple task with a deep neural network is the following:

1. Select a task f that the network has to learn.
2. Show the network some examples as pairs of points $(\mathbf{x}_i, f(\mathbf{x}_i))$.
3. With some optimization algorithm, find the value of network parameters (a's, $\mathbf{w}$'s and b's) that best fits the given examples.
4. Evaluate performance and use for the task

Our approach of study is to modify deep neural networks by adding a deterministic or random perturbation with the aim that the resulting models exhibit a forgetting behavior as similar as possible to that found in humans. As far as we know, no one has tried to study how one can make the resulting networks forget in a way similar to that exhibited in Ebbinghaus' experiments. Researchers have studied what happens if after learning one task the network is asked to learn another one. The results show, as expected, that eventually the new task takes over the previous one so the interesting problem is how long does it take for the new task to

take over and how many tasks can be subsequently learned without affecting the performance on the first one [19].

In the next subsection we present a single modification of neurons that makes the resulting deep neural network exhibit overall a similar forgetting behavior.

2.1 Making Regular Neurons Forget

In this subsection we will define in parallel the ***regular*** and a novel ***forgetting*** version of neural networks. We will refer to results being under the ***forgetting hypothesis*** when we want to emphasize that forgetting networks are considered. We will define them in terms of its fundamental units, which we call neurons, as defined next.

Definition 2.1 (Domain). We will define a ***domain*** in $\mathbb{R}^n$ as a Lebesgue measurable, connected set $\Omega \subseteq \mathbb{R}^n$.

Definition 2.2 (Neuron). Given a domain $\Omega \subseteq \mathbb{R}^n$, a ***neuron*** is a function $\eta : \Omega \to \mathbb{R}$ of the form

$$\eta(\boldsymbol{x}) = \sigma\big(\langle \boldsymbol{x}, \boldsymbol{w} \rangle + b\big) \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ are the ***weights***, $b \in \mathbb{R}$ the ***bias*** and $\sigma : \mathbb{R} \to \mathbb{R}$ the ***activation function***.

Definition 2.3 (Forgetting neuron). Given a domain $\Omega \subseteq \mathbb{R}^n$, a ***forgetting neuron*** is a function $\eta : \Omega \times [0, \infty] \to \mathbb{R}$ of the form

$$\eta(\boldsymbol{x}; t) = \varphi_a(t)\sigma\big(\langle \boldsymbol{x}, \boldsymbol{w}\varphi_w(t) \rangle + b\varphi_b(t)\big) \tag{2}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ are the ***weights***, $b \in \mathbb{R}$ the ***bias***, $\sigma : \mathbb{R} \to \mathbb{R}$ the ***activation function***, and φ_a, φ_w and φ_b are the ***forgetting functions***, that depend on a continuous parameter t (time).

2.2 Shallow Regular and Forgetting Networks

Definition 2.4 (Shallow network). Given a domain $\Omega \subseteq \mathbb{R}^n$ and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, a ***shallow network*** of N ***units*** is a linear combination of N neurons, i.e. it is a function $\Sigma : \Omega \to \mathbb{R}$ of the form:

$$\Sigma(\boldsymbol{x}) = \sum_{k=1}^{N} a_k \sigma\big(\langle \boldsymbol{x}, \boldsymbol{w}_k \rangle + b_k\big) \tag{3}$$

where $a_k \in \mathbb{R}$. We will denote the set of all shallow networks for a given σ, $\Omega \in \mathbb{R}^n$ and number of units N as

$$\mathcal{S}_{N,n}(\sigma, \Omega) \stackrel{\text{def}}{=} \left\{ \sum_{k=1}^{N} a_k \sigma\big(\langle \boldsymbol{x}, \boldsymbol{w}_k \rangle + b_k\big) \ : \ \boldsymbol{w}_k \in \mathbb{R}^n,\ a_k, b_k \in \mathbb{R} \right\} \tag{4}$$

and the set of all shallow networks as

$$\mathcal{S}_n(\sigma, \Omega) \stackrel{\text{def}}{=} \bigcup_{N=1}^{\infty} \mathcal{S}_{N,n}(\sigma, \Omega) \tag{5}$$

although we will usually drop the activation function and the domain and use $\mathcal{S}_{N,n}$ and $\mathcal{S}_n$ instead of $\mathcal{S}_{N,n}(\sigma, \Omega)$ and $\mathcal{S}_n(\sigma, \Omega)$.

Note that a shallow network of N units has $(n+2)N$ (trainable) parameters.

Definition 2.5 (Forgetting shallow network). Given a domain $\Omega \subseteq \mathbb{R}^n$, an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, forgetting functions $\varphi_a, \varphi_w, \varphi_b : [0, \infty) \to \mathbb{R}$ a ***forgetting shallow network*** of N ***units*** is a linear combination of N forgetting neurons. I.e., it is a function $\Sigma : \Omega \times [0, \infty) \to \mathbb{R}$ of the form:

$$\Sigma(\boldsymbol{x}; t) = \sum_{k=1}^{N} a_k \varphi_a(t) \sigma\big(\langle \boldsymbol{x}, \boldsymbol{w}_k \varphi_w(t) \rangle + b_k \varphi_b(t)\big) \tag{6}$$

where $a_k \in \mathbb{R}$.

In this definition we consider three individual forgetting functions, depending on the forgotten parameter involved (φ_a, φ_w and φ_b). Unless stated otherwise, along the paper we will suppose that forgetting occurs outside of the activation function (and therefore $\varphi_w(t) = \varphi_b(t) = 1$ for all t).

2.3 $\mathcal{G}$-functions

Definition 2.6 ($\mathcal{G}-$function). Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a connected directed acyclic graph (CDAG), being $\mathcal{V}$ and $\mathcal{E}$ the sets of vertices and edges respectively, with n source nodes and one sink node. For any $v \in \mathcal{V}$, let d_v be the number of in-edges of v. Consider that each $v \in \mathcal{V}$ has an associated function $f_v : \mathbb{R}^{d_v} \to \mathbb{R}$ (we call this the ***constituent function*** of v). Let $\Omega \subseteq \mathbb{R}^n$ be a domain. A $\mathcal{G}-$***function*** is a function $G : \Omega \to \mathbb{R}$ that is computed by the following rule:

- Each source node is a real variable input (since there are n nodes, the domain is in $\mathbb{R}^n$).
- In any other node v, each of the in-edges represents a real variable input, the node computes the result of its constituent function f_v, the result is thrown as an input to the vertices in each out-edges of v.
- The result of the whole network is the output of the only sink node.

Note that two different sets of constituent functions for the same CDAG $\mathcal{G}$ can give rise to the same $\mathcal{G}$-function.

Definition 2.7 (Internally $\mathcal{C}^k$ $\mathcal{G}$-function). A $\mathcal{G}$-function is said to be ***internally*** $\mathcal{C}^k$ if it admits some representation in which all its constituent functions are $\mathcal{C}^k$.
By analogy, we define ***internally*** $\mathcal{C}^k$ $\mathcal{G}$-functions.

Let us discuss some details about $\mathcal{G}$-functions. More specifically, we will give a justification for the concept of ***internally*** $\mathcal{C}^k$ functions and the issues associated.

The first thing to notice is that to extend theorems for shallow networks to their *deep* version we need the constituent functions to be representable by shallow networks, so we need them at least to be continuous. [2]

If no conditions are imposed to the constituent functions, all functions can be regarded as $\mathcal{G}$ functions, for any CDAG $\mathcal{G}$ with the right number of source nodes. This result will be formally stated later (Proposition 2.1).

The proof of this statement, relies on bijective functions between $\mathbb{R}$ and $\mathbb{R}^n$ and its inverses. It is a well known topological fact that those functions cannot be continuous.

The question that naturally arises is whether constituent functions can be limited to be continuous. This is not a new problem, and as far as we know there is no answer to that. An important related result is ***Kolmogorov-Arnold representation theorem*** ([5, 24]), which solved Hilbert's 13^{th} problem and states that any continuous multivariate function $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ has a decomposition of the form:

$$f(x_1, \ldots, x_n) = \sum_{q=1}^{2n} \Phi_q \left(\sum_{p=1}^{n} \varphi_{q,p}(x_p) \right) \qquad \Phi_q, \varphi_{q,p} \text{continuous} \tag{7}$$

[2] Or fulfill the more general condition of belonging to the set M defined in [29, Sec. 4].

This result requiring only one function Φ would directly state that any continuous function is in fact an internally continuous function. Although we have found no answer to the proposed question, we have found a paper from Giorsi and Poggio [18] stating that Kolmogorov's theorem is irrelevant because constituent functions are continuous but highly non-smooth, while there is another paper by Kůrková [25] satating that Kolmogorov theorem is indeed relevant.

If no condition is imposed to the constituent functions, then the following result holds:

Proposition 2.1. *For any CDAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and any function $f : \Omega \to \mathbb{R}$, there exists a set of constituent functions $\{h_v\}_{v\in\mathcal{V}}$ such that f is a $\mathcal{G}$ function with $\{h_v\}_{v\in\mathcal{V}}$ as constituent functions.*

This is an intuitive observation if one keeps in mind the existence of bijective functions between $\mathbb{R}$ and $\mathbb{R}^n$. We will not give explicit details on this result.

We have also found that if the constituent functions are forced to be twice continuously differentiable, then the analogous result does not hold. As a counterexample, we show here the case when $\mathcal{G}$ is a binary tree of 4 source nodes. This is our ***decreasing inference theorem***:

Theorem 2.2 (Decreasing Inference Theorem). *There exist CDAG's $\mathcal{G}$ with n source nodes and continuous functions $f \in \mathcal{C}(\mathbb{R}^n)$ such that f are not internally continuous $\mathcal{G}$-functions as long as the constituent functions are $\mathcal{C}^2$.*

PROOF.
With the above mentioned graph, a $\mathcal{G}$-function f has a decomposition of the form

$$f(x_1, x_2, x_3, x_4) = h\big(g_1(x_1, x_2), g_2(x_3, x_4)\big) \tag{8}$$

being its constituent functions $h, g_1, g_2 \in \mathcal{C}^2(\mathbb{R}^4)$. If we differentiate f with respect to x_1

$$\partial_{x_1} f(\boldsymbol{x}) = \partial_{y_1} h(g_1(x_1, x_2), g_2(x_3, x_4)) \cdot \partial_{x_1} g_1(x_1, x_2) \tag{9}$$

And differentiating again, this time with respect to x_3

$$\partial_{x_1 x_3} f(\boldsymbol{x}) = \partial_{y_1 y_2} h(g_1(x_1, x_2), g_2(x_3, x_4)) \cdot \partial_{x_3} g_2(x_3, x_4) \cdot \partial_{x_1} g_1(x_1, x_2) \tag{10}$$

By analogy, $\partial_{x_1 x_4} f(\boldsymbol{x})$ is

$$\partial_{x_1 x_4} f(\boldsymbol{x}) = \partial_{y_1 y_2} h(g_1(x_1, x_2), g_2(x_3, x_4)) \cdot \partial_{x_4} g_2(x_3, x_4) \cdot \partial_{x_1} g_1(x_1, x_2) \tag{11}$$

Now considering the quotient between (10) and (11)

$$\frac{\partial_{x_1 x_3} f(\boldsymbol{x})}{\partial_{x_1 x_4} f(\boldsymbol{x})} = \frac{\partial_{x_3} g_2(x_3, x_4)}{\partial_{x_4} g_2(x_3, x_4)} \tag{12}$$

Since the RHS does not depend on x_1, the LHS cannot depend on x_1. Therefore $\partial_{x_1}\left(\frac{\partial_{x_1x_3}f(\boldsymbol{x})}{\partial_{x_1x_4}f(\boldsymbol{x})}\right)=0$. Using derivation formulas, the numerator of $\partial_{x_1}\left(\frac{\partial_{x_1x_3}f(\boldsymbol{x})}{\partial_{x_1x_4}f(\boldsymbol{x})}\right)$ is

$$\big(\partial_{x_1x_3x_1}f(\boldsymbol{x})\big)\cdot\big(\partial_{x_1x_4}f(\boldsymbol{x})\big)-\big(\partial_{x_1x_4x_1}f(\boldsymbol{x})\big)\cdot\big(\partial_{x_1x_3}f(\boldsymbol{x})\big)=0 \tag{13}$$

Now, it gets easy to find a function that does not satisfy that condition. For example, set $f(x_1,x_2,x_3,x_4)=x_1x_3+x_1^2x_4$, it clearly does not fulfill the given condition. ■

A more general theory on this aspect is developed by Vitushkin (see [18, Th. 2.1]) who states the following theorem:

Theorem 2.3. *For any pair of natural numbers $k\geq 1$ and $n\geq 2$, there exist functions $f\in\mathcal{C}^k(I^n)$ that cannot be expressed as a superposition and composition of functions C^k of $n-1$ variables.*

2.4 Deep Regular and Forgetting Networks

Definition 2.8 (Layered graph). Let $\mathcal{G}=(\mathcal{V},\mathcal{E})$ be a CDAG. Let $V_1\subset\mathcal{V}$ be the set of source nodes, and for every integer $k>1$, let $V_k\subset\mathcal{V}$ be the set of nodes that have an in-edge from a node in V_{k-1}. $\mathcal{G}$ is a ***layered graph*** if for every pair of integers $i\neq j$, $V_i\cap V_j=\emptyset$

Note that if for some k, $V_k=\emptyset$, then it follows from the definition that for all $n\geq k$ $V_n=\emptyset$.

If $\mathcal{G}$ is layered, $\exists d\in\mathbb{N}$ such that $|V_d|=1$ and $V_{d+1}=\emptyset$. In this case, $\mathcal{V}=\bigsqcup_{i=1}^{d}V_i$. Each of these sets of vertices is called a ***layer***. V_1 is called the ***input layer***, V_d is called the ***output layer*** and the rest are called ***hidden layers***.

Definition 2.9 (Deep network). Let $\mathcal{G}=(\mathcal{V},\mathcal{E})$ be a CDAG with n source nodes and one sink node. Let $\Omega\subseteq\mathbb{R}^n$ be a domain. A ***$\mathcal{G}$-deep network*** is a $\mathcal{G}-$function $\Delta:\Omega\to\mathbb{R}$ with constituent functions being all shallow networks.

Definition 2.10 (Forgetting deep network). Let $\mathcal{G}=(\mathcal{V},\mathcal{E})$ be a CDAG with n source nodes and one sink node and $N=|\mathcal{V}|-n-1$ internal nodes. Let $\Omega\subseteq\mathbb{R}^n$ be a domain. A ***$\mathcal{G}$-forgetting deep network*** is a $\mathcal{G}-$function $\Delta:\Omega\times[0,\infty)\to\mathbb{R}$ with constituent functions being all forgetting shallow networks.

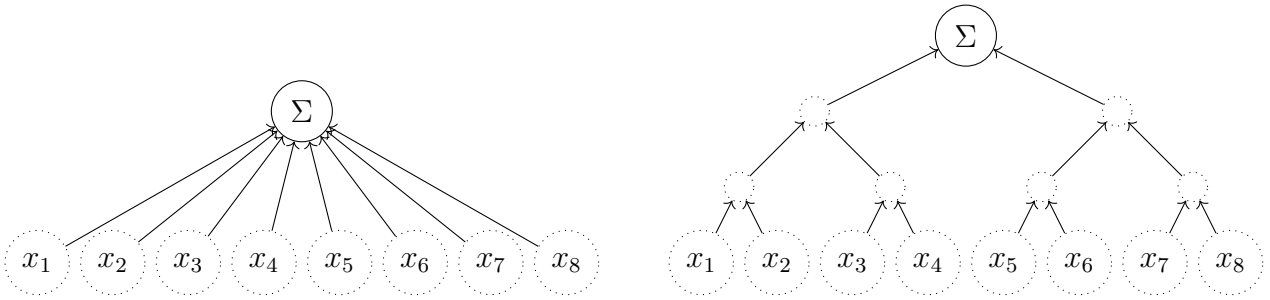

Figure 2: Shallow network *vs* binary-tree deep network.

In real life applications, one may be interested in functions with more than a single output. From a theoretical point of view, a function with many outputs can be regarded as many functions of a single output. This is why we will only consider single output functions in subsequent proofs.

One of the reasons for the interest in deep networks is that in most real world scenarios, functions have a $\mathcal{G}$-function structure [39, Appendix 2]. The reasoning comes from physics where it does not make sense that constituent functions are so pathological as the bijective functions between $\mathbb{R}$ and $\mathbb{R}^n$, and therefore our interest is focused in the internally continuous (or $\mathcal{C}^k$) case.

2.5 Reinforced Learning

In all the models we have considered that a single task is learned once and then it is gradually forgotten. In real situations, this only happens to the tasks that are not repeated (not reinforced).

Most interesting tasks are relearned again and again (like reading or driving) and we want to find a suitable way to model forgetting of these tasks.

To do so, we introduce a ***firing parameter*** to each neuron, so a forgetting neuron with reinforcement f is a function $\eta : \Omega \times [0, \infty] \times \mathbb{N} \to \mathbb{R}$ of the form:

$$\eta(\mathbf{x}; t, f) = \varphi(t, f)\sigma(\langle \mathbf{x}, \mathbf{w} \rangle + b) \tag{14}$$

where f is the firing parameter.

This firing parameter will be how many times the neuron has fired in the interval $[0, t]$. Typically, this can be defined in a proper way. For example, if $\sigma(x) = \max\{0, x\}$, a neuron is activated whenever $\sigma(x) > 0$. Or in the case of σ being the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, the neuron is fired when $\sigma(x) > 1/2$. We will assume that σ is such that this definition makes sense.

The firing parameter is considering reinforced learning, so φ should increase with f. Based on Ebbinghaus exponential forgetting, we postulate a forgetting function of the form:

$$\varphi(t, f) = e^{-\alpha t + \beta f} \tag{15}$$

for some $\alpha, \beta \in \mathbb{R}_+$. These auxiliary parameters model the speed at which the neuron forgets (α) and the effect of each reinforcement firing (β).

Following definitions on subsection 2.2, we can define a forgetting shallow network with reinforcement as a function $\Sigma : \Omega \times [0, \infty) \times \mathbb{N}^N \to \mathbb{R}$ of the form:

$$\Sigma(\mathbf{x}; t, f_1, \ldots, f_N) = \sum_{k=1}^{N} a_k \varphi(t, f_k) \sigma\big(\langle \mathbf{x}, \mathbf{w}_k \rangle + b_k\big) \tag{16}$$

And a forgetting deep network with reinforcement as a $\mathcal{G}-$function $\Delta : \Omega \times [0, \infty) \times \mathbb{N}^{N+1} \to \mathbb{R}$ with constituent functions being all forgetting shallow networks with reinforcement.

This definition may seem to add too many variables (in normal networks we had one variable in $\mathbb{R}^n$, and we have added one positive real variable and $N + 1$ discrete variables, we have added more variables than there were originally). In fact, for each training set, f_k will be a function of time (assuming our algorithm is deterministic). At the end of the day, we have a forgetting networks as in subsection 2.2 with a more complex forgetting function, that this time is different in each neuron.

Another possibility would be to consider the firing parameter to be a multi-integer for each neuron $\mathbf{f} = (f_1, \ldots, f_n)$.

The consequences of this approach will be part of our future research, but we want to highlight what happens with the results that will be presented in section 4.

In Theorem 4.1, Since the firing function is a discrete function, there is no firing, so $f_1 = \cdots = f_N = 0$. In that case, we recall exactly the same result for networks with reinforcement.

In Theorem 4.2 we distinguish two cases:

- If there is a finite number of reinforcements. In this case we get the same result of unavoidable forgetting, because from a certain point (the time when the last reinforcement is made) we can think the network as if it had no reinforcement at all.
- If there is arbitrary reinforcement. In this case the result is probably not true because if this reinforcement is made constantly to the same task, it is expected that the network does not forget that given task. What we do expect to be true is that if reinforcement is random and somehow uniform, we retrieve the same result in terms of expected value. This may be part of future work.

3 Forgetting Networks Estimate Functions in Sobolev Spaces

In this review section we introduce Sovolev spaces, which is the basic mathematical construct used in our forgetting proofs – they enable robust proofs that apply to any function learnable by deep neural networks. Users familiar with the related concepts can skip it. In the first section, we introduce normed function spaces, in the second Sovolev spaces (including the notions of weak derivative) and in the last basic function operators.

3.1 Normed Spaces

The goal of all learning algorithms is to start from a given function f and some information about this function, and then find a function f^* that is "similar enough" to the original function.

The statement becomes precise when we define a norm in the space of functions. In that case we can study the value $\|f - f^*\|$ to evaluate to what extent f^* is "similar enough" to f and if possible find an optimal within our representation abilities.

Definition 3.1 (Normed space). A ***norm*** in a vector space $\mathbb{X}$ is a function $\|\cdot\| : \mathbb{X} \to \mathbb{R}$ with the following properties:

i $\|x\| \geq 0$ for all $x \in \mathbb{X}$ and $\|x\| = 0 \iff x = 0$

ii $\|\lambda x\| = |\lambda|\|x\|$ for all $\lambda \in \mathbb{R}$ and all $x \in \mathbb{X}$

iii $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{X}$ (triangular inequality)

The pair $(\mathbb{X}, \|\cdot\|)$ is called a ***normed space***.

Definition 3.2 (p-norm, $\mathscr{L}^p$ spaces). Let $\Omega \in \mathbb{R}^n$ be a domain, $p \in [1, \infty)$. Consider $f : \Omega \to \mathbb{R}$, then its ***$p$-norm*** is (if it exists):

$$\|f\|_p \stackrel{\text{def}}{=} \left(\int_\Omega |f(\boldsymbol{x})|^p d\boldsymbol{x} \right)^{1/p} \tag{17}$$

This notion lets us define $\mathscr{L}^p$ spaces as:

$$\mathscr{L}^p(\Omega) \stackrel{\text{def}}{=} \{f : \Omega \to \mathbb{R} \ : \ \|f\|_p < \infty\} \tag{18}$$

For the case $p = \infty$, an analogous definition can be made, using the concept of essential supremum

$$\|f\|_\infty \stackrel{\text{def}}{=} \operatorname*{ess\,sup}_\Omega f = \inf\{a \in \mathbb{R} \ : \ f^{-1}(a, +\infty) \text{ is a set of measure zero in } A\} \tag{19}$$

The same definition can apply for $0 < p < 1$, but the resulting space is not a normed space. For $p \leq 0$, the problem is even worse because the norm is not defined for elementary functions like $f(x) = 0$ or $f(x) = x$.

Definition 3.3 ($\mathcal{C}^k$ spaces). Let $\Omega \subseteq \mathbb{R}^n$ and $k \in \mathbb{Z}_+$, we define the spaces

$$\mathcal{C}^k(\Omega) \stackrel{\text{def}}{=} \{f : \Omega \to \mathbb{R} \ : \ f \text{ has continuous partial derivatives up to order } k\} \tag{20}$$

$$\mathcal{C}^\infty(\Omega) \stackrel{\text{def}}{=} \bigcap_{k=1}^{\infty} \mathcal{C}^k(\Omega) \qquad ; \qquad \mathcal{C}(\Omega) \stackrel{\text{def}}{=} \{f : \Omega \to \mathbb{R} \ : \ f \text{ is continuous}\} \tag{21}$$

In all these spaces it is common the sup norm can be defined and it corresponds to the $\mathscr{L}^\infty$ norm for continuous functions.

Definition 3.4 (Lipschitz continuity). A function $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is ***Lipschitz continuous*** if there exists a constant C such that for all $x, y \in \Omega$ it is satisfied that $\|f(x) - f(y)\| \leq C\|x - y\|$.

If Ω is compact, this condition is stronger than continuity, but weaker than continuously differentiable. If $\Omega = \mathbb{R}^n$, there exist $\mathcal{C}^\infty$ functions that are not Lipschitz continuous, for example $f(x) = x^2$.

3.2 Sobolev Spaces

One technical detail to have in mind is that $\mathscr{L}^p$ as we have defined it is not rigorously a normed space, because there are non zero functions that have zero norm. To solve this, whenever two functions f, g satisfy $\|f - g\|_p = 0$, they will be considered the same function in $\mathscr{L}^p$.

This happens when $(f-g)^p = 0$ almost everywhere, that is equivalent to $f - g = 0$ almost everywhere. As a consequence if $\|f - g\|_p = 0$ for some p, then for all $q \in [1, \infty]$ $\|f - g\|_q = 0$. That is, whenever two functions f, g are considered the same in some $\mathscr{L}^p$ space, they are also considered the same in any other $\mathscr{L}^q$.

Since in $\mathscr{L}^p$ space the notion of the value of a function in a point has no meaning (a point is of measure zero), there is a priori no notion of derivative. This problem is solved by defining a suitable concept of weak derivative, that extends its classical version.

Consider two functions $F, \varphi \in \mathcal{C}^1(\mathbb{R})$, for some domain $I = [a, b] \subseteq \mathbb{R}$, it is well known that (integration by parts formula):

$$\int_a^b \frac{\partial F}{\partial x}\varphi = [F\varphi]_{x=a}^{x=b} - \int_a^b F\frac{\partial \varphi}{\partial x} \tag{22}$$

If we consider $\varphi \in \mathcal{C}_0^\infty(I)$, it is satisfied that $[F\varphi]_{x=a}^{x=b} = 0$, then the formula above becomes:

$$\int_a^b \frac{\partial F}{\partial x}\varphi = -\int_a^b F\frac{\partial \varphi}{\partial x} \tag{23}$$

If we consider all possible $\varphi \in \mathcal{C}_0^\infty(I)$, given $F \in \mathscr{L}^p(I)$, it can be shown that the formula:

$$\int_a^b \frac{\partial F}{\partial x}\varphi = -\int_a^b F\frac{\partial \varphi}{\partial x} \qquad \forall \varphi \in \mathcal{C}_0^\infty(I) \tag{24}$$

defines $\frac{\partial F}{\partial x}$ in the sense that it may or may not exist, but if it exists, $\frac{\partial F}{\partial x}$ is unique.

This can be generalized to n variables and derivatives of order k as follows.

Definition 3.5 (Weak derivative). Let $\Omega \subseteq \mathbb{R}^n$ be a domain, $F \in \mathscr{L}^p(\Omega)$. then the $\boldsymbol{k}$-th derivative of F can be defined (if it exists) as the only function satisfying:

$$\int_\Omega D^{\boldsymbol{k}} F \cdot \varphi = (-1)^{|\boldsymbol{k}|} \int_\Omega F \cdot D^{\boldsymbol{k}} \varphi \qquad \forall \varphi \in \mathcal{C}_0^\infty(\Omega) \tag{25}$$

where $\boldsymbol{k}$ is the multi-integer $\boldsymbol{k} = (k_1, \ldots, k_n)$, $|\boldsymbol{k}| = \sum_{i=1}^n k_i$ and $D^{\boldsymbol{k}} f = \dfrac{\partial^{|\boldsymbol{k}|} f}{\partial^{k_1} x_1 \cdots \partial^{k_n} x_n}$.

For a discussion on the existence and properties of weak derivatives, see [2, Ch. 3].

Example 1. Consider $f(x) = \begin{cases} 0 & \text{if } x \in \mathbb{Q} \\ \sin x & \text{if } x \notin \mathbb{Q} \end{cases}$ Since $\mathbb{Q}$ measure zero, this function is equal to $\tilde{f}(x) = \sin x$ in any $\mathscr{L}^p$, and as a differentiable function its derivative is $\cos x$.

Example 2. Another typical example is $f(x) = |x|$. This function has no classical derivative in $x = 0$. In this case it can be shown that the weak derivative is the sign function:

$$\sigma(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

This concept of weak derivative gives rise to the definition of Sobolev spaces, which are normed spaces with a certain number of (weak) derivatives.

Definition 3.6 (Sobolev norm). Let $f \in \mathscr{L}^p(\Omega)$. The ***Sobolev norm*** of a function $f \in \mathscr{L}^p(\Omega)$ defined (if all weak derivatives exist) as:

$$\|f\|_{p,m} \stackrel{\text{def}}{=} \sum_{0 \le |\boldsymbol{k}| \le m} \|D^{\boldsymbol{k}} f\|_p \tag{26}$$

Note that this sum has $\binom{n+m}{n}$ terms.

Definition 3.7 (Sobolev spaces). Given the space of functions $\mathscr{L}^p(\Omega)$, $\Omega \subseteq \mathbb{R}^n$. A ***Sobolev space*** in $\mathbb{R}^n$ with Sobolev norm $\|\cdot\|_{m,p}$ is the set of all functions $f : \mathbb{R}^n \to \mathbb{R}$ that can be weakly derived according to the multi integer $\boldsymbol{k}$ when $|\boldsymbol{k}| \le m$ and that have a Soolev norm smaller than one. Formally:

$$W_{p,m}^n(\Omega) \stackrel{\text{def}}{=} \{f \in \mathscr{L}^p(\Omega) \ : \ \|f\|_{p,m} \le 1\} \tag{27}$$

This ensures that for a function in the Sobolev space with m derivatives, this function must have derivatives up to order m in $\mathscr{L}^p$ and consequently derivatives do not get "too large".

3.3 Functional Operations within Sobolev Spaces: Convolution and Mollifiers

Definition 3.8 (Support, $\mathcal{C}_0^\infty(\Omega)$). Let f be a function defined in some domain $\Omega \subseteq \mathbb{R}^n$, the ***support*** of f is defined as:

$$\text{supp}(f) \stackrel{\text{def}}{=} \Omega \cap \text{closure}\{\boldsymbol{x} \in \Omega \ : \ f(\boldsymbol{x}) \neq 0\} \tag{28}$$

We say that a function f has compact support if its support is a compact subset of Ω. The set of k times differentiable functions will be denoted as:

$$\mathcal{C}_0^k(\Omega) \stackrel{\text{def}}{=} \{f \in \mathcal{C}^k(\Omega) \ : \ \text{supp}(f) \text{ is a compact set}\} \tag{29}$$

It will be of special interest the case of $k = \infty$.

Definition 3.9 (Ball). Let $\boldsymbol{x} \in \mathbb{R}^n$ and $r > 0$, we will denote $B(\boldsymbol{x}, r)$ the ***ball*** with center $\boldsymbol{x}$ and radius r. Formally

$$B(\boldsymbol{x}, r) \stackrel{\text{def}}{=} \{\boldsymbol{y} \in \mathbb{R}^n \ : \ |x - y| < r\} \tag{30}$$

Example 3. The typical example of a function with compact support is:

$$\eta(\boldsymbol{x}) = \begin{cases} ce^{\left(\frac{-1}{1-|\boldsymbol{x}|^2}\right)} & |\boldsymbol{x}| \leq 1 \quad (c \in \mathbb{R}) \\ 0 & |\boldsymbol{x}| > 1 \end{cases} \tag{31}$$

We can choose $c = \left(\int_{B(\mathbf{0},1)} e^{\left(\frac{-1}{1-|\boldsymbol{x}|^2}\right)} d\boldsymbol{x}\right)^{-1}$ so that the property $\int_{\mathbb{R}^n} \eta = 1$ is satisfied.

We can construct other examples from equation (31). Given $\varepsilon < 0$,

$$\eta_\varepsilon(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{\varepsilon^n} \eta\left(\frac{\boldsymbol{x}}{\varepsilon}\right) \tag{32}$$

has support $B(\mathbf{0}, \varepsilon)$ and also satisfies $\int_{\mathbb{R}^n} \eta_\varepsilon = 1$.

Definition 3.10 (Convolution). Let f, g be measurable functions in $\mathbb{R}^n$. The convolution of f and g is defined as

$$(f * g)(\boldsymbol{x}) \stackrel{\text{def}}{=} \int_{\mathbb{R}^n} f(\boldsymbol{y}) g(\boldsymbol{x} - \boldsymbol{y}) d\boldsymbol{y} \tag{33}$$

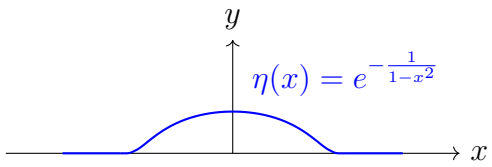

Figure 3: Graph of $\eta(x)$ for the unidimensional case

This integral may not exist depending on the shape of f and g. An interesting case, that we will use, is when $f \in \mathcal{C}(\mathbb{R})$ and $g \in \mathcal{C}_0^\infty(\mathbb{R})$. In that case $(f * g)(\boldsymbol{x})$ exists for all $\boldsymbol{x}$ and is infinitely differentiable by the following result

Lemma 3.1. *Given two measurable functions f, g*

i Convolution is commutative: $f * g = g * f$.

ii $f \in \mathcal{C}^j(\mathbb{R})$ and $g \in \mathcal{C}^k(\mathbb{R})$, then $f * g \in \mathcal{C}^{j+k}(\mathbb{R})$.

PROOF.
(i) This follows directly from a change of variable in the integral $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{y}$:

$$(f * g)(\boldsymbol{x}) = \int_{\mathbb{R}^n} f(\boldsymbol{y}) g(\boldsymbol{x} - \boldsymbol{y}) d\boldsymbol{y} \tag{34}$$

$$= \int_{\mathbb{R}^n} f(\boldsymbol{x} - \boldsymbol{z}) g(\boldsymbol{z}) d(\boldsymbol{x} - \boldsymbol{z}) = (g * f)(\boldsymbol{x}) \tag{35}$$

(ii) This follows from the derivative property: $\partial_{x_i}(f * g) = (\partial_{x_i} f) * g$:

$$\partial_{x_i}(f * g) = \partial_{x_i} \left(\int_{\mathbb{R}^n} f(\boldsymbol{y}) g(\boldsymbol{x} - \boldsymbol{y}) d\boldsymbol{y} \right) \tag{36}$$

$$\int_{\mathbb{R}^n} f(\boldsymbol{y}) \partial_{x_i}(g(\boldsymbol{x} - \boldsymbol{y})) d\boldsymbol{y} \qquad f \text{ is constant with respect to } \boldsymbol{x} \tag{37}$$

$$\int_{\mathbb{R}^n} f(\boldsymbol{y}) (\partial_{x_i} g)(\boldsymbol{x} - \boldsymbol{y}) d\boldsymbol{y} = f * \partial_{x_i} g \qquad \partial_{x_i}(\boldsymbol{x} - \boldsymbol{y}) = Id \tag{38}$$

■

In particular, if either f or g is infinitely differentiable, $f * g$ becomes also infinitely differentiable, regardless of the smoothness of the other.

The functions η_ε are called ***mollifiers*** because they have the following property (see Figure 4 for a numerical example)

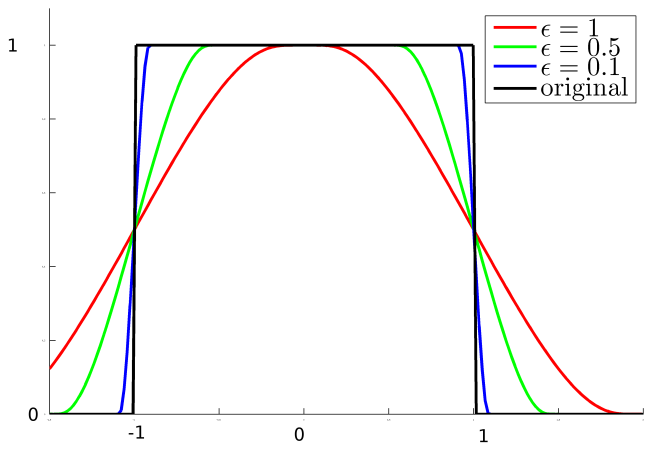

Figure 4: Graph of $H_\varepsilon(x)$ for different values of ε, being $H(x)$ the rectangle function: $H(x) = \begin{cases} 1 & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| > 1 \end{cases}$

Lemma 3.2. *Let $f \in \mathcal{C}(\Omega)$ and $f_\varepsilon = f * \eta_\varepsilon$. Then:*

1. $\text{supp}(f_\varepsilon) \subseteq \{\boldsymbol{x} \in \mathbb{R}^n \ : \ dist(\boldsymbol{x}, \text{supp}(f)) < \varepsilon\}$
2. $f_\varepsilon \in \mathcal{C}^\infty(\Omega)$
3. $f_\varepsilon \xrightarrow[\varepsilon \to 0]{} f$ *uniformly in each compact* $K \subseteq \mathbb{R}^n$.

PROOF.
The proof is made following [42, Lemma 7.1].

(i) Let $S = \text{supp}\, f$. It suffices to prove that if $\text{dist}(x, S) \geq \varepsilon$, then $f_\varepsilon(\boldsymbol{x}) = 0$.

$f_\varepsilon(\boldsymbol{x}) = \int_{\mathbb{R}^n} \eta_\varepsilon(\boldsymbol{z}) f(\boldsymbol{x} - \boldsymbol{z}) d\boldsymbol{z} = \int_{B(\boldsymbol{0},\varepsilon)} \eta_\varepsilon(\boldsymbol{z}) f(\boldsymbol{x} - \boldsymbol{z}) d\boldsymbol{z}$ because $\text{supp}(\eta_\varepsilon) = B(\boldsymbol{0}, \varepsilon)$.

If $|\boldsymbol{z}| < \varepsilon$ and $\text{dist}(\boldsymbol{x}, S) \geq \varepsilon$, triangular inequality states that $\text{dist}(\boldsymbol{x} - \boldsymbol{z}, S) \geq \text{dist}(\boldsymbol{x}, S) - \text{dist}(\boldsymbol{z}, S) > 0$, then $\boldsymbol{x} - \boldsymbol{z} \notin \text{supp}\, f$, so $f(\boldsymbol{x} - \boldsymbol{z}) = 0$. From the definition of f_ε as an integral with $f(\boldsymbol{x} - \boldsymbol{z})$ as a factor, directly follows $f_\varepsilon(\boldsymbol{x}) = 0$.

(ii) This is a direct consequence of Lemma 3.1 because $\eta_\varepsilon \in \mathcal{C}_0^\infty$.

(iii) Since $\int_{\mathbb{R}^n} \eta_\varepsilon = 1$, we can write

$$f_\varepsilon(\boldsymbol{x}) - f(\boldsymbol{x}) = \int_{\mathbb{R}^n} \eta_\epsilon(\boldsymbol{z}) \left[f(\boldsymbol{x} - \boldsymbol{z}) - f(\boldsymbol{x}) \right] d\boldsymbol{z} \tag{39}$$

In this form we can see that for any $\boldsymbol{x} \in \Omega$, we have that

$$|f_\varepsilon(\boldsymbol{x}) - f(\boldsymbol{x})| \leq \sup_{B(\boldsymbol{0},\varepsilon)} |f(\boldsymbol{x} - \boldsymbol{z}) - f(\boldsymbol{x})| \tag{40}$$

If we consider $K \subseteq \Omega$ compact, then f is uniformly continuous in K, therefore $\sup_{B(\mathbf{0},\varepsilon)} |f(\boldsymbol{x}-\boldsymbol{z}) - f(\boldsymbol{x})| \xrightarrow[\varepsilon\to 0]{} 0$ uniformly for $\boldsymbol{x} \in K$. This uniform convergence together with (40) gives the result that $f_\varepsilon - f \xrightarrow[\varepsilon\to 0]{} 0$ uniformly in K. ■

3.4 Asymptotic Notation

Asymptotic notation is used in this paper and in many of the references. For the sake of completeness and clearness, formal definitions are included next:

Definition 3.11 ($\mathcal{O}, o, \Omega, \omega, \Theta, \theta$). Let $g : \mathbb{R} \to \mathbb{R}$ be a real function and $a \in \mathbb{R} \cup \{\pm\infty\}$. We define the following sets of functions:

- $\mathcal{O}(g(x)) \stackrel{\text{def}}{=} \{f : \mathbb{R} \to \mathbb{R} \ : \ \lim_{x\to a} \frac{f(x)}{g(x)} < \infty\}$
- $o(g(x)) \stackrel{\text{def}}{=} \{f : \mathbb{R} \to \mathbb{R} \ : \ \lim_{x\to a} \frac{f(x)}{g(x)} = 0\}$
- $\Omega(g(x)) \stackrel{\text{def}}{=} \{f : \mathbb{R} \to \mathbb{R} \ : \ \lim_{x\to a} \frac{f(x)}{g(x)} > 0\}$
- $\omega(g(x)) \stackrel{\text{def}}{=} \{f : \mathbb{R} \to \mathbb{R} \ : \ \lim_{x\to a} \frac{f(x)}{g(x)} = \infty\}$
- $\Theta(g(x)) \stackrel{\text{def}}{=} \mathcal{O}(g(x)) \cap \Omega(g(x))$
- $\theta(g(x)) \stackrel{\text{def}}{=} \{f : \mathbb{R} \to \mathbb{R} \ : \ \lim_{x\to a} \frac{f(x)}{g(x)} = 1\}$

These definitions are also applied for natural functions $g : \mathbb{N} \to \mathbb{N}$ when $a = \infty$.

The mathematical sentence $f \in \mathcal{O}(g)$ is read as "f ***is*** $\mathcal{O}(g)$", and is typically written as $f = \mathcal{O}(g)$, an abuse of notation that is widely used and emphasizes the fact that we are not interested in f but in its asymptotic behavior. This applies for all the classes defined above.

Typically the value a is omitted and it is understood to be either $a = 0$ or $a = \infty$ depending on the context.

4 Basic Behavior of Forgetting: Theorem of Non-Instantaneous Forgetting and Theorem of Universal Forgetting

In this section we will prove two results necessary for a realistic approach to the modeling of forgetting by neural networks: that it is instantaneous and unavoidable in the long term. We remark that for these results to hold our proof only assumes that the forgetting function $\varphi(t)$ is continuous with respect to t and that $\lim_{t\to\infty} \varphi(t) = 0$. In fact, both are direct results of continuity, as we show in the detailed proof in the next subsections.

4.1 Forgetting Is Not Instantaneous

In this subsection we prove that forgetting cannot happen instantaneously under our model. Formally:

Theorem 4.1. ***(Theorem of non-instantaneous forgetting)*** *Let Δ be a forgetting deep network defined over the compact domain Ω with activation function $\sigma \in \mathcal{C}^1(\mathbb{R})$. For all $\varepsilon > 0$, there exists $\delta t > 0$ such that:*

$$\|\Delta(\,\cdot\,;\delta t) - \Delta(\,\cdot\,;0)\| < \varepsilon \tag{41}$$

when the norm used is the sup norm.

PROOF.
For a given $\boldsymbol{x} \in \Omega$, we define the function $D_{\boldsymbol{x}}(t) \stackrel{\text{def}}{=} \Delta(\boldsymbol{x}; t)$.

In that case, $D_{\boldsymbol{x}}(t)$ is continuous with respect to t because it is the composition of continuous functions (because the only dependence on t is on the functions $\varphi(t)$, which are clearly continuous by hypothesis). If $\varphi(t)$ could jump discontinuously to 0, forgetting would too.

Given ε, continuity of $D_{\boldsymbol{x}}(t)$ implies $\forall \boldsymbol{x} \in \Omega$, $\exists \delta t$, that depends on $\boldsymbol{x}$ and ε, such that

$$\forall t \in (-\delta t, \delta t) \qquad |D_{\boldsymbol{x}}(0) - D_{\boldsymbol{x}}(\delta t)| < \varepsilon \tag{42}$$

if we look now at $D_{\boldsymbol{x}}(t)$ as a function of $\boldsymbol{x}$ (with fixed t), since σ is continuously differentialbe, $D_x(t)$ is Lipshitz continuous with respect to $\boldsymbol{x}$. This means that for all $t \in \mathbb{R}$, there exists $C_t > 0$ such that $|D_{\boldsymbol{x}}(t) - D_{\boldsymbol{y}}(t)| \leq C_t \|\boldsymbol{x} - \boldsymbol{y}\|$.

With these tools, we will prove that given $\varepsilon > 0$, for all $\boldsymbol{x} \in \Omega$, there exists $r_{\boldsymbol{x}} > 0$ and δt, such that for all $\boldsymbol{y} \in \Omega$ and $t \in \mathbb{R}$

$$(\|x - y\| < r_{\boldsymbol{x}} \text{ and } |t| < \delta t) \implies |D_{\boldsymbol{y}}(0) - D_{\boldsymbol{y}}(t)| < \varepsilon \tag{43}$$

Continuity of $D_{\boldsymbol{x}}(t)$ with respect to t states that, considering $\varepsilon/3$, , there exists δt fulfilling equation (42). Let C_t be the Lipshitz constant of $D_{\boldsymbol{x}}(t)$ as stated before, we define $C = \max_{t \in [-\delta t, +\delta t]} \{C_t\}$. It is satisfied:

$$\forall t \in [-\delta t, +\delta t] \qquad |D_{\boldsymbol{x}}(t) - D_{\boldsymbol{y}}(t)| \leq C \|\boldsymbol{x} - \boldsymbol{y}\| \tag{44}$$

Now if we define $r_{\boldsymbol{x}} = \frac{\varepsilon}{3C}$, we have, for all $\boldsymbol{y}, t$ such that $\|\boldsymbol{x} - \boldsymbol{y}\| < r_{\boldsymbol{x}}$ and $|t| < \delta t$:

$$|D_{\boldsymbol{y}}(0) - D_{\boldsymbol{y}}(t)| = |D_{\boldsymbol{y}}(0) - D_{\boldsymbol{x}}(0) + D_{\boldsymbol{x}}(0) - D_{\boldsymbol{x}}(t) + D_{\boldsymbol{x}}(t) - D_{\boldsymbol{y}}(t)| \tag{45}$$

$$\leq |D_{\boldsymbol{y}}(0) - D_{\boldsymbol{x}}(0)| + |D_{\boldsymbol{x}}(0) - D_{\boldsymbol{x}}(t)| + |D_{\boldsymbol{x}}(t) - D_{\boldsymbol{y}}(t)| \quad \text{Triangular inequality} \tag{46}$$

$$\leq C\|\boldsymbol{x} - \boldsymbol{y}\| + \varepsilon/3 + C\|\boldsymbol{x} - \boldsymbol{y}\| \quad \text{Equations (44) and (42)} \tag{47}$$

$$\leq C r_{\boldsymbol{x}} + \varepsilon/3 + C r_{\boldsymbol{x}} = C\frac{\varepsilon}{3C} + \frac{\varepsilon}{3} + C\frac{\varepsilon}{3C} = \varepsilon \quad \text{Definition of } r_{\boldsymbol{x}} \tag{48}$$

This proves equation (43). Now we will use the fact that Ω is compact. Consider the open cover of Ω:

$$\Omega = \bigcup_{\boldsymbol{x}\in\Omega} B(\boldsymbol{x}, r_{\boldsymbol{x}}) \tag{49}$$

where $B(\boldsymbol{x}, r_{\boldsymbol{x}})$ is the open ball centered in $\boldsymbol{x}$ and radius $r_{\boldsymbol{x}}$ as defined before. Since Ω is compact, there exists $N \in \mathbb{N}$ and a finite number of elements $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ such that

$$\Omega = \bigcup_{i=1}^{N} B(\boldsymbol{x}_i, r_{\boldsymbol{x}_i}) \tag{50}$$

For each $\boldsymbol{x}_i$ consider $(\delta t)_{\boldsymbol{x}_i}$ that fulfills equation (43). We define [3] $\delta t = \min_{i=1:N}\{(\delta t)_{\boldsymbol{x}_i}\}$.

The proof is finished because given $\varepsilon > 0$, δt as we have just defined fulfills

$$|\Delta_{\boldsymbol{x}}(0) - \Delta_{\boldsymbol{x}}(t)| < \varepsilon \qquad \forall \boldsymbol{x} \in \Omega \text{ if } |t| < \delta t \tag{51}$$

Since we are using the *sup* norm, this is directly what we want to prove. ■

A good analogy for this is that forgetting can be seen as *diming the light*. In other words, if you are in a place and clearly see everything, after a small diming of the light, you can see clearly most of what you could see before.

4.2 Forgetting Is Unavoidable

In this section we prove that, if there is not reinforcement of a particular task, under our model everything will be eventually forgotten. Formally:

Theorem 4.2. ***(Theorem of universal forgetting)*** *Let Δ be a forgetting deep network defined over the compact domain Ω with activation function $\sigma \in \mathcal{C}(\mathbb{R})$. Then*

$$\lim_{t\to\infty} \|\Delta(\,\cdot\,; t)\| = 0 \tag{52}$$

when the norm used is the sup norm.

PROOF.
This result is a direct consequence of continuity of φ and σ. We follow by induction on N the number of nodes of the graph. The base case is $n = 2$, in which case the only dependence on t if $\varphi(t)$ multiplying the network, so clearly the limit is zero.

Let $\mathcal{G}$ be the graph associated to Δ and let f be the constituent function of the only sink node.

$$f(x_1, \ldots, x_{d_s}; t) = \sum_{i=1}^{N_s} a_k \varphi(t) \sigma\big(\langle \boldsymbol{x}, \boldsymbol{w}_k\rangle + b_k\big) \tag{53}$$

[3] In this step the importance of Ω being compact becomes clear. If it wasn't, the minimum δt may be zero.

For each in-edge of the sink node, consider v_k the vertex the edge is coming from. Since the degree of the sink node is d_s, there are d_s different vertices. For each of these vertices, consider $\mathcal{G}_k$ the maximal subgraph of $\mathcal{G}$ such that $\mathcal{G}_k$ is a CDAG with v_k as its only sink node. If we consider the same constituent functions as in $\mathcal{G}$ this constructions gives rise to deep networks $\Delta_1, \ldots, \Delta_{d_s}$ each one of them with less nodes than the original network. We will apply the induction hypothesis there. For any $\boldsymbol{x} \in \mathbb{R}^n$, let us denote the vector $\boldsymbol{X}(\boldsymbol{x}; t) = (\Delta_1(\boldsymbol{x}_1; t), \ldots, \Delta_{d_s}(\boldsymbol{x}_{d_s}; t))$. With this notation:

$$\Delta(\boldsymbol{x}; t) = \sum_{k=1}^{N_s} a_k \varphi(t) \sigma\Big(\langle \boldsymbol{X}, \boldsymbol{w}_k \rangle + b_k\Big) \tag{54}$$

The induction hypothesis implies that $\lim_{t \to 0} \boldsymbol{X}(\boldsymbol{x}; t) = \boldsymbol{0}$. Taking limits when $t \to 0$ in $\Delta(\boldsymbol{x}; t)$:

$$\begin{aligned}
\lim_{t \to 0} \Delta(\boldsymbol{x}; t) &= \sum_{k=1}^{N_s} a_k \lim_{t \to 0} \varphi(t) \lim_{t \to 0} \sigma\Big(\langle \boldsymbol{X}, \boldsymbol{w}_k \rangle + b_k\Big) && \text{Linearity of limits, if both exist} \\
&= \sum_{k=1}^{N_s} a_k \lim_{t \to 0} \varphi(t) \sigma\Big(\langle \lim_{t \to 0} \boldsymbol{X}, \boldsymbol{w}_k \rangle + b_k\Big) && \text{Continuity of } \sigma \\
&= \sum_{k=1}^{N_s} a_k \lim_{t \to 0} \varphi(t) \sigma(b_k) && \text{Induction hypothesis} \\
&= \sum_{k=1}^{N_s} a_k \cdot 0 \cdot \sigma(b_k) = 0 && \Big(\lim_{t \to 0} \varphi(t) = 0\Big)
\end{aligned}$$

■

5 Higher Frequencies Are Forgotten Faster: Universality Theorems and The Forgetting "Center of Mass Theorem"

A fundamental question that has been answered about neural networks is its representation potential: are they able to approximate any function to any precision?

The proof depends on the graph $\mathcal{G}$ associated with the network and the activation function σ. The main result of this section, obtained in [29, Th. 2.1] is that a shallow network with activation function satisfying certain week hypothesis (the result even works for many functions with essential discontinuities) are universal if and only if the activation function σ is not a polynomial. This is a very powerful result and its complete proof uses advanced analysis, so we will prove the particular case of σ being continuous, which we consider illustrative enough for our purpose, and a wide enough result, since most learning algorithms use continuous activation functions.

As part of the proof we will show that forgetting deep neural networks forget "higher frequencies" faster, essentially "generalizing" the examples, this is what we call the *center of mass theorem.*

5.1 Center of Mass and Universality Theorems

Definition 5.1 (Density property). Let $\mathbb{X}$ be a space of functions with some domain $\Omega \subseteq \mathbb{R}^n$, and let $\mathcal{F} \subseteq \mathbb{X}$ be a family of functions in this space. We say that $\mathcal{F}$ is ***universal*** or ***dense*** if for every $g \in \mathbb{X}$ and for every compact $K \subseteq \Omega$, there exists a sequence of functions $\{f_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$, such that

$$\lim_{i\to\infty} \|g - f_i\|_{\mathscr{L}^\infty(K)} = 0 \tag{55}$$

We have defined density with the *sup* norm. An analogous definition can be stated with any other p-norm, but we have chosen to fix the norm for the sake of simplicity.

In particular, latter in the section we will comment how our theorems generalize to p-norms.

The two main theorems in this section are the following:

Theorem 5.1 (Universality theorem for $\sigma \in \mathcal{C}(\mathbb{R})$). *Shallow networks with an arbitrary number of units and activation function $\sigma \in \mathcal{C}(\mathbb{R})$ are universal in $\mathcal{C}(\mathbb{R}^n)$ $\iff$ σ is not a polynomial.*

Theorem 5.2 (Center of mass theorem). *Considering shallow networks with the **forgetting hypothesis** to be that only weights (w's) are forgotten (i.e. $\varphi_a(t) = \varphi_b(t) = 1$ for all t) and a polynomial target function; then each monomial is forgotten as $x^k \to (\varphi(t)\cdot x)^k$, so high degree elements of the polynomial are forgotten much faster than small degree elements.*

5.2 Proof of the Center of Mass and Universality Theorems

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{m} = (m_1, \ldots, m_n)$. be vectors in $\mathbb{R}^n$. We use the following notation:

$$\boldsymbol{x}^{\boldsymbol{m}} = x_1^{m_1} \cdots x_n^{m_n} \qquad ; \qquad |\boldsymbol{m}| = m_1 + \cdots + m_n \tag{56}$$

Let H_k^n be the set of homogeneous polynomials of n coordinates and degree k and P_k^n be the set of polynomials of degree at most k (for P_k^n we include the case $k = \infty$). I.e.:

$$H_k^n = \left\{ \sum_{|\boldsymbol{m}|=k} a_{\boldsymbol{m}} \boldsymbol{x}^{\boldsymbol{m}} \right\} \qquad P_k^n = \bigcup_{i=0}^{k} H_i^n \tag{57}$$

Note that H_k^n is a vector space of dimension $\binom{n-1+k}{k}$

The main result of this section is Theorem 5.7 and its corollaries. To arrive there, we will need some previous results:

Lemma 5.3. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be an infinitely differentiable function. Then the following conditions are equivalent:*

i σ is not a polynomial.

ii There exists $x \in \mathbb{R}$ such that for all $n \in \mathbb{N}$ $\sigma^{(n}(x) \neq 0$.

The proof of this lemma is explained in appendix A, subsection A.1.

Lemma 5.4. *Let $\Omega \subseteq \mathbb{R}^n$, let $L(\Omega) = \bigcup_{\boldsymbol{a}\in\Omega} \operatorname{span}\{\boldsymbol{a}\}$. If the only polynomial in P_∞^n that vanishes in $L(\Omega)$ is the trivial one, then the set*

$$\mathscr{M}(\Omega) = \operatorname{span}\{g(\boldsymbol{a}\cdot\boldsymbol{x}) \ : \ \boldsymbol{a} \in \Omega,\ g \in \mathcal{C}(\mathbb{R})\} \tag{58}$$

is dense in $\mathcal{C}(\mathbb{R}^n)$.

PROOF.

We will prove that under the stated hypothesis, for any k, $H_k^n \subseteq \mathscr{M}(\Omega)$. A generalized version of this proof can be found in [30, Theorem 2.1]. Applying Stone-Weierstrass's theorem [27, Ch. 3] one can complete the proof.

First we prove that for any $\boldsymbol{d} \in L(\Omega)$, $(\boldsymbol{d}\cdot\boldsymbol{x})^k \in \mathscr{M}(\Omega)$.

To do so, consider $\boldsymbol{d} \in L(\Omega)$, by the definition of $L(\Omega)$, $\boldsymbol{d} \in \operatorname{span}(\boldsymbol{a})$ for some $\boldsymbol{a} \in \Omega$, so there exists $y \in \mathbb{R}$ such that $\boldsymbol{d} = y\boldsymbol{a}$. So the function $(\boldsymbol{d}\cdot\boldsymbol{x})^k$ can be written as: $(\boldsymbol{d}\cdot\boldsymbol{x})^k = (y\boldsymbol{a}\boldsymbol{x})^k = g(\boldsymbol{x}\cdot\boldsymbol{a}) \in \mathscr{M}(\Omega)$, with $g(x) = (yx)^k$.

Now consider the dual space of H_k^n, defined as

$$H_k^{n*} = \{\sigma : H_k^n \to \mathbb{R} \ : \ \sigma \text{ is a linear function}\} \tag{59}$$

Since H_k^n is a finite vector space, H_k^{n*} is a finite vector space of the same dimension. A basis of this space is $V = \{D^{\boldsymbol{m}} \ : \ |\boldsymbol{m}| = k\}$ where

$$\begin{aligned} D : H_k^n &\to \mathbb{R} \\ p(\boldsymbol{x}) &\mapsto D^{\boldsymbol{m}}p(\boldsymbol{x}) = \frac{\partial^{|\boldsymbol{m}|}p(\boldsymbol{x})}{\partial x_1^{m_1}\cdots\partial x_n^{m_n}} \end{aligned}$$

It is a basis because it has the right number of elements and they are mutually independent since:

$$D^{\boldsymbol{m}}\boldsymbol{x}^{\boldsymbol{m'}} = \begin{cases} 0 & \text{if } \boldsymbol{m} \neq \boldsymbol{m'} \\ m_1!\cdots m_n! & \text{if } \boldsymbol{m} = \boldsymbol{m'} \end{cases} \tag{60}$$

Since V is a basis of H_k^{n*}, any element can be written as a linear combination of elements in V. Equivalently, any linear function that maps H_k^n to $\mathbb{R}$ can be written in terms of a polynomial $q \in H_k^n$ as

$$\begin{aligned} f_q : H_k^n &\to \mathbb{R} \\ p &\mapsto q(D)p \end{aligned}$$

We want to study how this functions f_q act on functions of the form $(\boldsymbol{d}\cdot\boldsymbol{x})^k$. We first study the case of q being a monomial, i.e. $q(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{m}}$ for some $\boldsymbol{m} \in \mathbb{Z}_+^n$, $|\boldsymbol{m}| = k$. Applying derivative properties:

$$q(D)(\boldsymbol{d}\cdot\boldsymbol{x})^k \;=\; D^{\boldsymbol{m}}(d_1x_1 + \cdots + d_nx_n)^k \;=\; kd_1D^{(m_1-1,\ldots,m_n)}(\boldsymbol{d}\cdot\boldsymbol{x})^{k-1} \;=\; \cdots \;=\; k!q(\boldsymbol{d}) \quad (61)$$

Using linearity, the previous result generalizes to all $q \in H_k^n$. With this result, we want to study what happens when we apply f_q functions to the linear subspace of polynomials

$$W = \text{span}\{(\boldsymbol{d}\cdot\boldsymbol{x})^k \in :\; d \in L(\Omega)\} \subseteq H_k^n \quad (62)$$

If some $f_q \in H_k^{n*}$ annihilates all polynomials in W, in particular it annihilates all polynomials of the form $(\boldsymbol{d}\cdot\boldsymbol{x})^k$ for all $\boldsymbol{d} \in L(\Omega)$. Since $f_q(\boldsymbol{d}\cdot\boldsymbol{x})^k = k!q(\boldsymbol{d})$, this means that q vanishes in $L(\Omega)$. By hypothesis, $q = 0$ (as a polynomial). In terms of linear algebra, W is a subspace of H_k^n with the property that any linear function $\sigma : H_n^k \to \mathbb{R}$ is trivial if and only if the restriction $\sigma\big|_W : W \to \mathbb{R}$ is trivial. This implies $W = H_k^n$. The proof is finished observing $H_k^n \subseteq W \subseteq \mathscr{M}(\Omega)$. ■

The next corollary follows immediately from the proof of this lemma and will be useful for another result (Theorem 6.1).

Corollary 5.5. *Let $r = dim\ H_k^n = \binom{n-1+k}{k}$ and $s = dim\ P_k^n = \binom{n+k}{k}$. Then there exist $\{\boldsymbol{a}^i\}_{i=1:r} \subseteq \mathbb{R}^n$, $\{\boldsymbol{b}^i\}_{i=1:s} \subseteq \mathbb{R}^n$, $\{f_i\}_{i=1:r} \subseteq H_k^n$ and $\{g_i\}_{i=1:s} \subseteq P_k^n$ such that*

$$H_k^n = \left\{\sum_{i=1}^r f_i(\boldsymbol{a}^i\cdot\boldsymbol{x}) \;:\; f_i \in H_k^1\right\} \qquad P_k^n = \left\{\sum_{i=1}^s g_i(\boldsymbol{b}^i\cdot\boldsymbol{x}) \;:\; g_i \in P_k^1\right\} \quad (63)$$

PROOF.
Consider the case $\Omega = \mathbb{R}^n$ of the previous lemma. For each j, we have that

$$H_j^n = \text{span}\{(\boldsymbol{d}\cdot\boldsymbol{x})^j \in :\; d \in \mathbb{R}^n\} \quad (64)$$

As a consequence, there exist $\{\boldsymbol{a}^i\}_{i=1:r} \subseteq \mathbb{R}^n$ such that $\{(\boldsymbol{a}^i\cdot\boldsymbol{x})^j\}_{i=1:r}$ is a basis of H_j^n. This is equivalent to

$$H_j^n = \left\{\sum_{i=1}^r g_i(\boldsymbol{a}^i\cdot\boldsymbol{x}) \;:\; g_i \in H_j^1\ \forall i\right\} \quad (65)$$

The analogous version for non homogeneous polynomials follows directly because $P_k^n = \bigcup_{j=1}^k H_j^n$ ■

Lemma 5.6. *If $\mathcal{S}_1(\sigma, \mathbb{R})$ is dense in $\mathcal{C}(\mathbb{R})$, then $\mathcal{S}_n(\sigma, \mathbb{R}^n)$ is dense in $\mathcal{C}(\mathbb{R}^n)$.*

PROOF.
Let $g \in \mathcal{C}(\mathbb{R}^n)$ and $K \subseteq \mathbb{R}^n$ compact. Lemma 5.4 states that $\mathscr{M} = \text{span}\{f(\boldsymbol{a} \cdot \boldsymbol{x}) \ : \ \boldsymbol{a} \in \mathbb{R}^n, \ f \in \mathcal{C}(\mathbb{R})\}$ is dense in $\mathcal{C}(K)$. Thus given $\varepsilon > 0$, there exists $k \in \mathbb{N}$ and a sequence of functions $\{f_i\}_{i=1:k} \subseteq \mathcal{C}(\mathbb{R})$ and a sequence of vectors $\{\boldsymbol{a}^i\}_{i=1:k} \subseteq \mathbb{R}^n$ such that for all $\boldsymbol{x} \in K$:

$$\left| g(\boldsymbol{x}) - \sum_{i=1}^{k} f_i(\boldsymbol{a}^i \cdot \boldsymbol{x}) \right| < \frac{\varepsilon}{2} \tag{66}$$

Since K is compact, for all $i = 1 : k$ there exists a finite interval $[\alpha_i, \beta_i]$ such that

$$\{\boldsymbol{a}^i \cdot \boldsymbol{x} \ : \ \boldsymbol{x} \in K\} \subseteq [\alpha_i, \beta_i] \tag{67}$$

Because $\mathcal{S}_1$ is dense in $[\alpha_i, \beta_i]$, there exists $m_1 \in \mathbb{N}$ and constants $c_{ij}, w_{ij}, \theta_{ij}$ such that for all $y \in [\alpha_i, \beta_i]$:

$$\left| f_i(y) - \sum_{j=1}^{m_i} c_{ij} \sigma \left(w_{ij} y + \theta_{ij} \right) \right| < \frac{\varepsilon}{2k} \tag{68}$$

Applying both inequalities yields, for all $\boldsymbol{x} \in K$:

$$\left| g(\boldsymbol{x}) - \sum_{i=1}^{k} \sum_{j=1}^{m_i} c_{ij} \sigma \left(w_{ij} y + \theta_{ij} \right) \right| < \varepsilon \tag{69}$$

■

Theorem 5.7 (Universality Theorem for $\sigma \in \mathcal{C}^\infty(\mathbb{R})$). *Shallow networks with an arbitrary number of units and activation function $\sigma \in \mathcal{C}^\infty(\mathbb{R})$ are universal in $\mathcal{C}(\mathbb{R}^n)$ $\iff$ σ is not a polynomial.*

PROOF.

$\boxed{\Longleftarrow}$

On behalf of Lemma 5.6, we only need to do the proof for $n = 1$.

Consider w and b fixed. For any $h > 0$, we have that

$$\frac{\sigma \left(x(w + h) + b \right) - \sigma \left(xw + b \right)}{h} \in \mathcal{S}_1 \tag{70}$$

Hence, the limit for $h \to 0$ is in its closure $\overline{\mathcal{S}_1}$. This limit is $\frac{\partial}{\partial w} \big(\sigma \left(xw + b \right) \big)$. We can apply an analogous argument using the same argument to prove that any k-th derivative is in $\overline{\mathcal{S}_1}$ (see Figure 5 for a visual explanation). Applying differentiation rules we have that:

$$\frac{\partial^k}{\partial w^k} (\sigma(xw + b)) = x^k \sigma^{(k}(xw + b) \in \overline{\mathcal{S}_1} \tag{71}$$

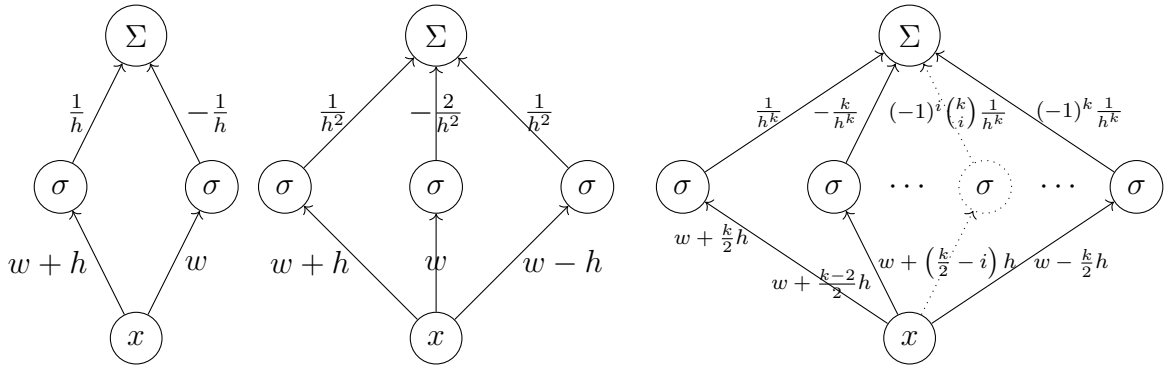

Figure 5: Illustration of how the k-th derivative is obtained using networks with k neurons (assuming $\sigma^{(k}(0) \neq 0$). The input is represented by $\textcircled{x}$, each arrow represents a product and $\textcircled{\sigma}$ and $\textcircled{\Sigma}$ represent the activation and the sum functions respectively. So for example, the first picture represents $\frac{1}{h}\sigma\Big(x(w+h)\Big) - \frac{1}{h}\sigma(xw)$, which is the approximation for the first derivative $\frac{\partial}{\partial w}\sigma(xw) = x\sigma'(xw)$. We have explicitly added w, but in most cases we can suppose $w = 0$ (as explained in the proof of Theorem 5.7.

Since σ is not a polynomial, Lemma 5.3 guarantees that there exists a number $\tilde{b}$ such that for any k, $\sigma^{(k}(\tilde{b}) \neq 0$. Taking $w = 0$ and $b = \tilde{b}$, we have that for any integer k, the monomial $x_k \in \overline{\Sigma_1}$. As a consequence, $\overline{\mathcal{S}_n}$ contains all polynomials.

By Stone-Weierstrass's theorem [27, Ch. 3], its closure contains all continuous functions.

$\boxed{\Longrightarrow}$

We prove by contradiction. Suppose σ is a polynomial of degree k. Then $\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle + b)$ is a polynomial degree at most k for any $\boldsymbol{w}$ and b. Thus, the family Σ_n is contained in the family of polynomials of degree at most k, which is not dense in $\mathcal{C}(\mathbb{R}^n)$. This is a classical density result, in appendix A, subsec:universalityShallow we give the details for the sake of completeness,for the case $n = 1$.

■

The proof of the analogous theorem demanding the activation function only to be continuous (Theorem 5.1) is explained in appendix A, subsection A.4.

Corollary 5.8 (Density of deep networks in internally continuous $\mathcal{G}$-functions). *Let $\mathcal{G}$ be a CDAG. Then $\mathcal{G}$-Deep networks with an arbitrary number of units and activation function $\sigma \in \mathcal{C}^\infty$ are universal in $\mathcal{C}(\mathbb{R}^n) \cap \{$internally continuous $\mathcal{G}$-functions$\}$ $\iff$ σ is not a polynomial.*

The proof is simply done by applying the theorem to each constituent function. We omit more explicit details.

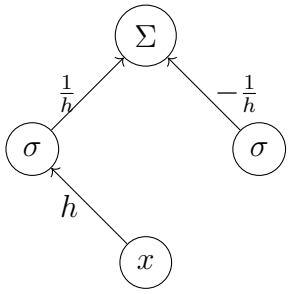

Figure 6: A special and interesting case happens when $b = 0$ and $\sigma'(0) \neq 0$. In this case, setting $w = 0$, the monomial x can be computed as $1/\sigma'(0)$ times the simple network of this figure. We consider this case specially interesting because it needs one less connection. This could be biologically more efficient. Moreover, this little network depends only on h, the parameter that gives the precision, so the same network could be used in many places of the brain where the basic monomial x is needed.

We want to note that in real applications, for this result to apply, you need to guess the compositional form of the function you want to approximate beforehand. As we already commented in section 2, we conjecture that this is a critical point, since there are functions that do not have a given compositional form when internal functions are forced to be continuous (as it is the case for neural networks).

If this conjecture is true, then given a complex task (a function) with a concrete compositional structure (corresponding to a distribution of basic tasks), the ones which are in the lower levels behave independently to the ones on the higher levels (see Figure 7).

We also want to highlight the following fact, that we will use in the proof of Theorem 6.1.

Corollary 5.9. *Let H_k^1 be the set of single variable polynomials of degree most k. Then $H_k^1 \subseteq \overline{\mathcal{S}_{k+1,1}}$. And in general, for polynomials of n variables and degree k, $H_k^n \subseteq \overline{S_{s\left(\frac{k+n}{n}\right)^n,n}}$, where $s = \dim H_k^n = \binom{n+k-1}{k}$.*

PROOF.
In the proof of Theorem 5.7 we have seen that $P_k^1 \subseteq \overline{\mathcal{S}_1}$ by saying that a multiple of the monomial x^k can be seen as a k-th derivative of $\sigma(wx + b)$ with respect to w for some $b \in \mathbb{R}$. This k-th derivative can be approximated by functions in $\mathcal{S}_1$. In particular, using the finite differences approach (for a visual representation see Figure 5, and for a developed theory, see [11, Ch. 3]) to approximate such derivative, a derivative of order k with can be approximated with $k + 1$ evaluations of the function. In our context this means using $k + 1$ units.

For the multivariate case, we observe that in order to approximate the monomial $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ we need at most $\prod_{i=1}^n (\alpha_i + 1)$ units. This number follows from the subsequent reasoning: for each variable x_i, the monomial $x_i^{\alpha_i}$ can be approximated by a suitable network of $\alpha_i + 1$

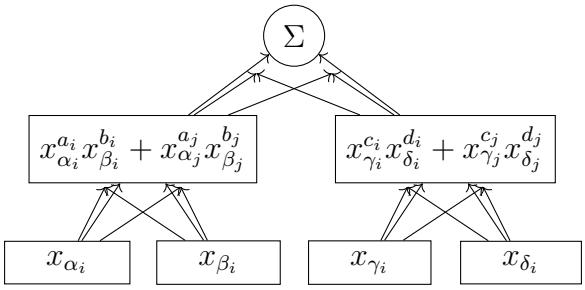

Figure 7: Features of very different kind (made of different sets of variables x_α, x_γ for example) behave independently and can only be aggregated in higher levels of the network, corresponding to more abstract concepts.

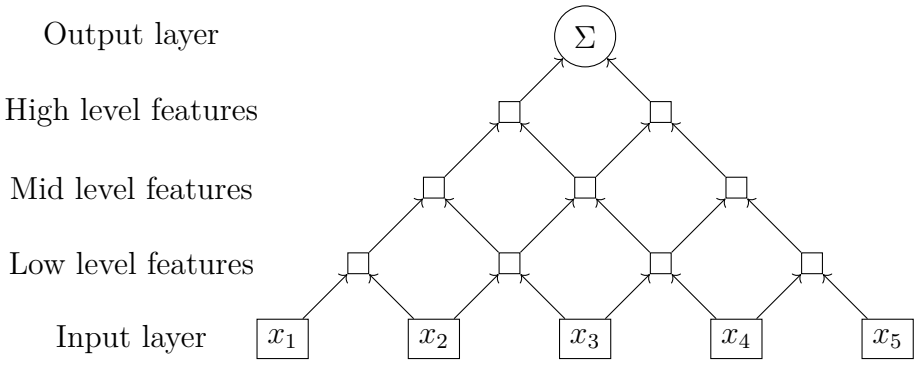

Figure 8: Conceptual distribution of features into different levels. In the case of vision, for example, these levels may coincide with the three stages of vision described by David Marr [32].

units. If we do it for x_1, we can apply the finite differences method cited in the univariate case for the network obtained, so using $(\alpha_1+1)\cdot(\alpha_2+1)$ units the monomial $x_1^{\alpha_1}\cdot x_2^{\alpha_2}$ can be obtained. The same argument can be extended to n variables for any n.

Now we know $\sum_{i=1}^{n}(\alpha_i+1)=k+n$, so using the inequality between arithmetic and geometric mean, we can find a bound for the number of units required for each monomial:

$$\left(\prod_{i=1}^{n}(\alpha_i+1)\right)^{1/n}\leq\frac{k+n}{n}\implies\prod_{i=1}^{n}(\alpha_i+1)\leq\left(\frac{k+n}{n}\right)^{n}\tag{72}$$

Since a base of H_k^n has exactly s monomials, that can be each approximated by networks with at most $\left(\frac{k+n}{n}\right)^n$ units, we have that $H_k^n\subseteq\overline{S_{s\left(\frac{k+n}{n}\right)^n,n}}$ as stated. ■

In this corollary the first of the bounds is optimum, because k-th derivatives cannot be approximated by finite differences with less than $k+1$ points. In contrast, the bound for the multivariate case is highly non optimal for two reasons. First one, the bound for $\prod_{i=1}^{n}(\alpha_i+1)$ is pretty strong, and in fact the equality happens only in a small number of cases, and only for k's that are multiple of n. The second reason is because we are not sure that our method for computing the multivariate derivative $\frac{\partial^k}{\partial^{\alpha_1}x_1\cdots\partial^{\alpha_n}x_n}$ uses the minimum possible number of neurons.

The method of the proof gives an easy rule to compute the number of neurons needed to approximate a given polynomial. As an example, the number of neurons needed to compute a simple polynomial as $xy+xy^2$ would be $2\cdot2+2\cdot3=10$.

Obviously, if one wants to generate all polynomials in n variables of degree at most k, the previous corollary can be applied adding the result homogeneous polynomials of degree for $i=0:k$. We think this number can be improved, and in fact, for the case of single variable polynomials, we have that $P_k^1\subseteq\mathcal{S}_{2k+1,1}$. This is because the k-th derivative is obtained as the limit when $h\to0$ of $\sum_{i=0}^{k}(-1)^i\binom{k}{i}\frac{1}{h^k}\sigma\left(x\left(w+h\left(\frac{k}{2}-i\right)\right)+b\right)$. The key observation is that, if k is even, the value of the parameters needed to compute the k-th derivative contains the values for all even numbers smaller than k, and if k is odd, contains all the values for odd numbers smaller than k. Then the linear combination of units can be properly arranged to compute any polynomial (see example below).

Example 4. We want to build a network that approximates $f(x)=x^3+3x^2+x+1$. Lemma 5.3 states that there exists $\tilde{b}$ such that $\sigma^{(k}(\tilde{b})\neq0$. In this example, for the sake of simplicity, we suppose that $\tilde{b}=0$.

1. First we approximate each of the monomials:
   (a) $x^3\approx\frac{1}{\sigma^{(3}(0)h^3}\Big(\sigma(3xh/2)-3\sigma(xh/2)+3\sigma(-xh/2)-\sigma(-3xh/2)\Big)$
   (b) $3x^2\approx\frac{3}{\sigma^{(2}(0)h^2}\Big(\sigma(xh)+2\sigma(0)+\sigma(-xh)\Big)$
   (c) $2x\approx\frac{2}{\sigma'(0)h}\Big(\sigma(xh/2)-\sigma(-xh/2)\Big)$
   (d) $1\approx\frac{1}{\sigma(0)}\Big(\sigma(0)\Big)$

2. Write the linear combination of the monomials

$$x^3+3x^2+x+1 \approx \frac{1}{\sigma^{(3}(0)h^3}\Big(\sigma(3xh/2)-3\sigma(xh/2)+3\sigma(-xh/2)-\sigma(-3xh/2)\Big)+ \\ +\frac{3}{\sigma^{(2}(0)h^2}\Big(\sigma(xh)+2\sigma(0)+\sigma(-xh)\Big)+ \\ +\frac{2}{\sigma'(0)h}\Big(\sigma(xh/2)-\sigma(-xh/2)\Big)+\frac{1}{\sigma(0)}\Big(\sigma(0)\Big)$$

and rearrange the terms

$$x^3+3x^2+x+1 \approx \frac{1}{\sigma^{(3}(0)h^3}\sigma(3xh/2)+\frac{3}{\sigma^{(2}(0)h^2}\sigma(xh)+\left(-3\frac{1}{\sigma^{(3}(0)h^3}+\frac{2}{\sigma'(0)h}\right)\sigma(xh/2)+ \\ +\left(2\frac{3}{\sigma^{(2}(0)h^2}+\frac{1}{\sigma(0)}\right)\sigma(0)+\left(3\frac{1}{\sigma^{(3}(0)h^3}-\frac{2}{\sigma'(0)h}\right)\sigma(-xh/2)+ \\ +\frac{-3}{\sigma^{(2}(0)h^2}\sigma(-xh)+\frac{-1}{\sigma^{(3}(0)h^3}\sigma(-3xh/2)$$

The following figure provides a graphical representation

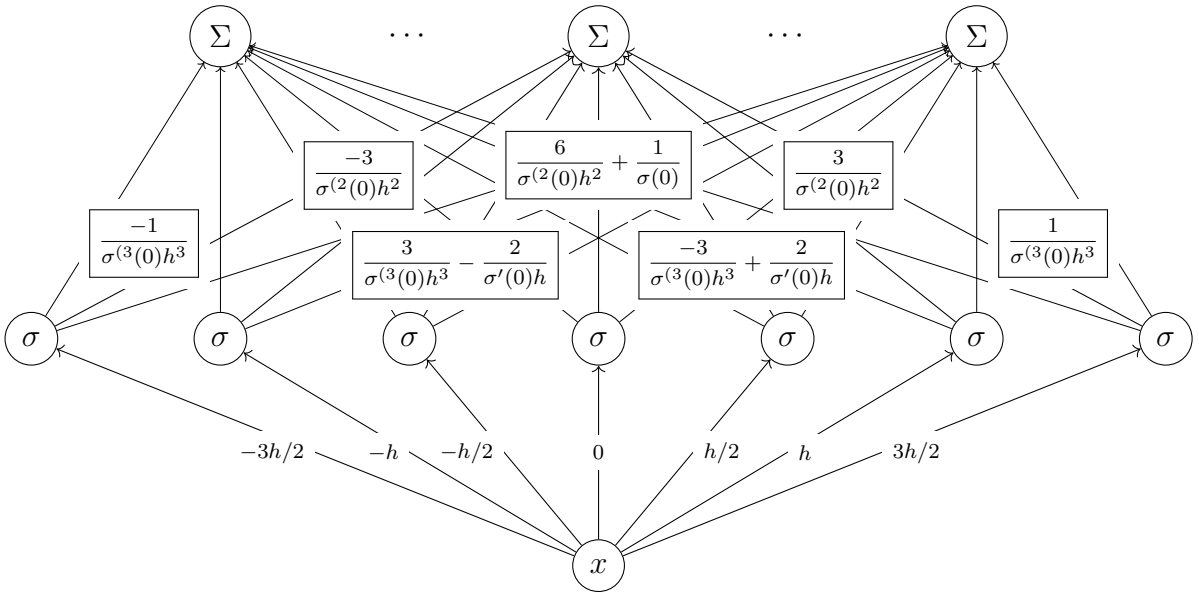

Figure 9: Approximation of $f(x)=x^3+3x^2+x+1$ using 7 units instead of 10. The other Σ's represent any other 3-order polynomial that can be generated with some combination of the same set of 7 neurons. As long as σ is such that $\sigma^{(k}(0)\neq 0$ for every k the above 7 "basis" can generate any 3-order polynomial. In general, with $2N+1$ neurons we can generate any polynomial of order at most N. Similar constructs could be made for other function representations such as radial basis functions ([10], [44], [3]) or Gröbner polynomials ([9]).

PROOF.
(Of Theorem 5.2) The proof of Theorem 5.7 is based on the fact that polynomials can be used to approximate any continuous function and than we can obtain any polynomial as a limit of shallow networks (see equation (71)). If we consider the case of forgetting networks, substituting x by $x\varphi_w(t)$, where $\varphi_w(t)$ is the forgetting function of the weights [4], the polynomial we obtain is

$$(x\varphi_w(t))^k \sigma^{(k}(xw+b) \tag{73}$$

The effect of $\varphi_w(t)$ increases with the degree of the resulting polynomial. ■

This theorem is directly related with forgetting high frequencies faster than lower ones because from the perspective of polynomials, high frequencies are associated with high degree polynomials.

Representations Other than Polynomials To prove the universality theorem we have given a way to approximate a given function with a neural network: you build the networks approximating each monomial, and with them you can build any polynomial, that will approximate arbitrarily well your target function. We think other representations may be useful, for example Göbner basis [9].

5.3 Generalization to Other Norms

Definition 5.1 can be generalized to a different topology or convergence criterion saying that a family $\mathcal{F}$ is dense in $\mathbb{X}$ if for every $g \in \mathbb{X}$, there exist $\{f_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$ such that $f_i \xrightarrow[i\to\infty]{} g$ for a given convergence criterion. Our previous definitions and theorems are for the topology of uniform convergence on compacta.

We want to note that these results can be extended to any $\mathscr{L}^p(\Omega)$ space over a compact domain Ω because $\mathcal{C}(\Omega)$ is dense in $\mathscr{L}^p(\Omega)$ for any $p \in [1,\infty]$ and Ω compact.

5.4 Forgetting High Frequencies in Deep Networks

Theorem 5.2 is stated only for shallow networks. In deep networks the modeling of forgetting may be more complex since it can depend on the graph. Each layer may be forgotten in a different way, and even each neuron may have different $\varphi_a, \varphi_b, \varphi_w$. However, if we consider the forgetting hypothesis with $\varphi_b(t) = \varphi_w(t) = 1$ for all t, in the second layer we have a phenomenon analogous to the forgetting hypothesis with $\varphi_a(t) = \varphi_b(t) = 1$ for all t in a single layer, so Theorem 5.2 directly applies. These kind of arguments may be applied to gain intuition in how some specific networks forget, regarding high and low frequencies.

[4] This is equivalent to substitute w by $w\varphi_w(t)$ as stated in the new forgetting hypothesis.

6 More Neurons Do Not Delay Forgetting: Curse of Forgetting Theorem

From the previous section we know that both shallow and deep networks are universal, meaning they can approximate any function to an arbitrary degree of accuracy if sufficient neurons are added.

The problem of this result is that it is true when there is no limit to the number of units in the network. In a real world situation, the number of units is constrained by computational limitations.

In this section we present two fundamental results, one regarding shallow networks and another showing that deep networks can avoid the curse of dimensionality.

6.1 Curse of Dimensionality and Forgetting

Theorem 6.1 (Curse of dimensionality). *Let $(\mathbb{X}, \|\cdot\|_p)$ be a normed space with $p \in [1, \infty]$. Let $\sigma \in \mathcal{C}^\infty(\mathbb{R})$ and not a polynomial. For $f \in W^n_{\infty,m}([-1,1]^n)$, the complexity of the shallow networks that provide accuracy at least ϵ is*

$$N = \mathcal{O}\left(\epsilon^{-n/m}\right) \tag{74}$$

and it is the best possible among all reasonable methods of approximation. By reasonable we mean the ones described after the definition of non-linear N-width.

We present also the impact of ***forgetting*** on this theorem in the form of another theorem.

Theorem 6.2 (The curse of forgetting). *Consider the forgetting version of the network, with forgetting function $\varphi(t)$. If the network Σ_f approximates f with accuracy ε, then the corresponding network after time t, $\Sigma_f(\cdot, t)$ approximates f with accuracy at least*

$$\tilde{\varepsilon}(t) = \varepsilon(2 - \varphi(t)) + (1 - \varphi(t)) \tag{75}$$

and this is the best possible upper bound.

To study this size-constrained problem the following definition will be useful :

Definition 6.1 (Approximation error). Let $(\mathbb{X}, \|\cdot\|)$ be a normed space, $W \subseteq \mathbb{X}$ and $f \in \mathbb{X}$. The ***best approximation error*** of f in W is

$$E(f; W; \|\cdot\|) \stackrel{\text{def}}{=} \inf_{g \in W} \|f - g\| \tag{76}$$

Let $V \subseteq \mathbb{X}$. The ***worst case approximation error*** of V in W is

$$E(V; W; \|\cdot\|) \stackrel{\text{def}}{=} \sup_{f \in V} \inf_{g \in W} \|f - g\| \tag{77}$$

We will usually drop the norm when it is understood from the context which norm is considered. In that case we will write $E(f; W)$ instead of $E(f; W; \|\cdot\|)$.

In this section we will focus in the following problem: given $\epsilon > 0$, given $W \subseteq \mathbb{X}$, and $\mathcal{S}_{N,n} \subseteq \mathbb{X}$ the set of neural networks with less or equal than N units, which is the minimum N such that

$$E\left(W; \mathcal{S}_{N,n}\right) < \epsilon \tag{78}$$

We will use asymptotic notation in this section. For example, if $\text{dist}(f, \mathcal{S}_{N,n}) = \mathcal{O}\left(N^{-\gamma}\right)$ for some $\gamma > 0$, then a network with complexity $N = \mathcal{O}\left(\epsilon^{-\frac{1}{\gamma}}\right)$ is enough to guarantee an approximation of accuracy at least ϵ.

The results we present are based on [39] and rely on the fact that the space W of target functions is somehow controlled. We will think of Sobolev spaces as in Definition 3.7 with the sup norm and over $\Omega = [-1, 1]^n$, so we will generally write W_m^n instead of $W_{\infty,m}^n([-1, 1]^n)$.

To prove both theorems we will need four basic lemmas that we will cover next. First, a classical result from approximation theory:

Lemma 6.3. *Given P_k^n the space of polynomials of degree at most k in n variables and W_m^n the Sobolev space as defined in Definition 3.7 with the sup norm, there exists a constant C such that the following inequality holds:*

$$E(\mathrm{W}_m^n; P_k^n) \leq Ck^{-m} \tag{79}$$

Since the proof is long and is not specially relevant, the reader is referred tosubsection A.3.

To prove that the complexity given is the best possible, we need to define the concepts of ***Bernstein N-width*** and ***continuous non-linear N-width*** (see [1] and [37, Sec. 6])

Definition 6.2 (Bernstein N-width). Given $\mathbb{X}$ a normed linear space and $K \subseteq \mathbb{X}$ a compact subset of it, the ***Bernstein N-width*** is:

$$b_N(K; X) \stackrel{\text{def}}{=} \sup_{X_{N+1}} \sup\{\lambda \,:\, \lambda S(X_{d+1}) \subseteq K\} \tag{80}$$

Where X_{N+1} is any $(N+1)$-dimensional subspace of $\mathbb{X}$ and $S(X_{N+1})$ is the unit ball of X_{N+1}.

Definition 6.3 (Continuous non-linear N-width). Given $\mathbb{X}$ a normed linear space and $K \subseteq \mathbb{X}$ a compact subset of it. Let $P_N : K \to \mathbb{R}^N$ be a continuous function and let $M_N : \mathbb{R}^N \to \mathbb{X}$ be any function. For each such P_N and M_N set

$$E(K; P_N, M_N; \mathbb{X}) \stackrel{\text{def}}{=} \sup_{f \in K} \|f - M_N(P_N(f))\| \tag{81}$$

and now define the ***continuous non-linear N-width*** as

$$h_N(K; \mathbb{X}) \stackrel{\text{def}}{=} \inf_{P_N, M_N} E(K; P_N, M_N; \mathbb{X}) \tag{82}$$

The idea behind this definition is the following:

- A learning algorithm can be regarded as a function $\Lambda : K \to \mathbb{X}$, because given a target function $f \in K$, returns its approximating neural network, which is a function in $\mathbb{X}$.
- This function Λ can be factorized into two functions $\Lambda : K \xrightarrow{P_N} \mathbb{R}^N \xrightarrow{M_N} \mathbb{X}$. P_N maps every function to a set of parameters, and given a set of parameters, M_N returns its corresponding neural network as a function in $\mathbb{X}$.
- Given an approximating algorithm (i.e. given P_N and M_N), $E(K; P_N, M_N; \mathbb{X})$ is the worst case error of the considered algorithm.
- Given K and $\mathbb{X}$, $h_N(K; \mathbb{X})$ represents the minimum $E(K; P_N, M_N; \mathbb{X})$ over all possible algorithms (P_N, M_N).

So our $E(K, \mathcal{S}_{n,N})$ is equal to $h_{(n+2)N}(K, \mathbb{X})$ when we restrict learning algorithms (P_N, M_N) to be P_N continuous and M_N the exact one described in the definition of shallow networks.

Lemma 6.4. *For any normed space $\mathbb{X}$ and $K \subseteq \mathbb{X}$ compact*

$$h_N(K; \mathbb{X}) \geq b_N(K; \mathbb{X}) \tag{83}$$

PROOF.
Let $P_N : K \to \mathbb{R}^N$ be a continuous function. Set

$$\tilde{P_N}(f) = P_N(f) - P_N(-f) \tag{84}$$

Thus $\tilde{P}_N : K \to \mathbb{R}^N$ is an odd continuous function. Given X_{N+1} an $(N+1)$-dimensional subspace of $\mathbb{X}$ and $\lambda > 0$ such that $\lambda S(X_{N+1}) \subseteq K$, then $\tilde{P}_N\Big|_{\partial(\lambda S(X_{N+1}))}$ is an odd continuous function from the boundary of an $(N+1)$-dimensional ball to $\mathbb{R}^N$. By Borsuk-Ulam theorem, there exists an $f^* \in \partial(\lambda S(X_{N+1}))$ (in particular $\|f\| = \lambda$) for which $\tilde{P}_N(f^*) = 0$. As a consequence, for any function $M_N : \mathbb{R}^N \to \mathbb{X}$

$$2f^* = [f^* - M_N(P_N(f^*))] - [-f^* - M_N(P_N(-f^*))] \tag{85}$$

and therefore

$$\max\{\|f^* - M_N(P_N(f^*))\|,\ \|-f^* - M_N(P_N(-f^*))\|\} \geq \|f^*\| = \lambda \tag{86}$$

Since both f^* and $-f^*$ are in K, this implies that $E(K; P_N, M_N; \mathbb{X}) \geq \lambda$. Since this inequality is valid for any choice of P_N and M_N and $\lambda \leq b_N(K; \mathbb{X})$, we have that $h_N(K; \mathbb{X}) \geq b_N(K; \mathbb{X})$ and the proof is done. ■

Lemma 6.5. *With the notation previously defined, there exists a constant C such that*

$$b_N(W_m^n; c) \geq CN^{-m/n} \tag{87}$$

PROOF.
Since b_N is a supremum, it suffices to prove that there exists a constant C and an $(N+1)$-dimensional linear space X_{N+1} such that $CN^{-m/n}S(X_{N+1}) \subseteq W_m^n$.

Let φ be any nonzero function in $\mathcal{C}^\infty(\mathbb{R}^n)$ with $\operatorname{supp}\varphi \subseteq [-1,1]^n$. For given n, m, we can choose φ satisfying $\|D^{\mathbf{k}}\varphi\| \leq 1$ for $|k| \leq m$. For $l > 0$ and $\mathbf{j} \in (2\mathbb{Z})^n$, set

$$\varphi_{\mathbf{j},l}(x_1, \ldots, x_n) = \varphi(x_1 l - j_1, \ldots, x_n l - j_n) \tag{88}$$

The support of $\varphi_{\mathbf{j},l}$ lies in $\prod_{i=1}^n [(j_i - 1)/l, (j_i + 1)/l]$. Since we are working with sup norm, we have that:

$$\|\varphi_{\mathbf{j},l}\| = \|\varphi\| \qquad \|D^{\mathbf{k}}\varphi_{\mathbf{j},l}\| = l^{|\mathbf{k}|}\|D^{\mathbf{k}}\varphi\| \tag{89}$$

For any fixed l, we observe that for different $\mathbf{j} \in (2\mathbb{Z})^n$, the supports of $\varphi_{\mathbf{j},l}$ are disjoint. Therefore, for any linear combination we have:

$$\left\|\sum_{\mathbf{j}} c_{\mathbf{j}}\varphi_{\mathbf{j},l}\right\| = \|c\|_\infty\|\varphi\| \qquad \left\|D^{\mathbf{k}}\left(\sum_{\mathbf{j}} c_{\mathbf{j}}\varphi_{\mathbf{j},l}\right)\right\| = l^{|\mathbf{k}|}\|c\|_\infty\|D^{\mathbf{k}}\varphi\| \tag{90}$$

where $\|c\|_\infty = \max_{\mathbf{j}} |c_{\mathbf{j}}|$.

$$\left\|\sum_{\mathbf{j}} c_{\mathbf{j}}\varphi_{\mathbf{j},l}\right\|_{m,\infty} = \sum_{0\leq|\mathbf{k}|\leq m}\left\|D^{\mathbf{k}}\left(\sum_{\mathbf{j}} c_{\mathbf{j}}\varphi_{\mathbf{j},l}\right)\right\| = \quad \text{(See Definition 3.6)} \tag{91}$$

$$= \sum_{0\leq|\mathbf{k}|\leq m} l^{|\mathbf{k}|}\|c\|_\infty\|D^{\mathbf{k}}\varphi\| \leq \|c\|_\infty \sum_{0\leq|\mathbf{k}|\leq m} l^m = \quad (\|D^{\mathbf{k}}\varphi\| \leq 1 \text{ and } l^k \leq l^m) \tag{92}$$

$$= \|c\|_\infty\binom{n+m}{m}l^m = \binom{n+m}{m}l^m\left\|\sum_{\mathbf{j}} c_{\mathbf{j}}\varphi_{\mathbf{j},l}\right\| \quad \text{The sum has } \tbinom{n+m}{n} \text{ terms.} \tag{93}$$

So for l large, the linear space generated by those $\varphi_{\mathbf{j},l}$ whose support lie totally in $[-1,1]^n$ is a linear space of dimension of the order of l^n with the property that

$$\left\|\sum_{\mathbf{j}} c_{\mathbf{j}}\varphi_{\mathbf{j},l}\right\|_{\infty} \leq 1 \implies Cl^{-m}\left\|\sum_{\mathbf{j}} c_{\mathbf{j}}\varphi_{\mathbf{j},l}\right\|_{m,\infty} \leq 1 \tag{94}$$

for some constant C independent of l. This implies that $b_N(W_m^n;\mathcal{C}([-1,1]^n)) \geq Cl^{-m}$ where $N \approx l^n$. Thus we have proved the desired result

$$b_N(W_m^n;\mathcal{C}([-1,1]^n)) \geq CN^{-m/n} \tag{95}$$

■

PROOF.
(Of Theorem 6.1) The first proof of this result was given in [33]. We will proceed with a slightly different approach, following [37], we will prove that $E(\mathrm{W}_m^n;\mathcal{S}_{N,n}) \leq CN^{-m/n}$ for a suitable constant C independent of N.

We begin recalling Corollary 5.5 and Corollary 5.9, that together tell us

$$P_k^n = \left\{\sum_{i=1}^{s} g_i(\mathbf{a}^i \cdot \mathbf{x}) \ : \ g_i \in \overline{\mathcal{S}_{k+1,n}}\ \forall i\right\} \tag{96}$$

where $s = \dim P_k^n = \binom{n+k}{k}$.

From this result it follows directly that:

$$P_k^n \subseteq \overline{\mathcal{S}_{s(k+1),n}} \tag{97}$$

Set $N = s(k+1)$. Then there exists a constant C' independent of N such that

$$\begin{aligned} E(\mathrm{W}_m^n;\mathcal{S}_{N,n}) &= E(\mathrm{W}_m^n;\overline{\mathcal{S}_{N,n}}) && && (98)\\ &\leq E(\mathrm{W}_m^n;P_k^n) && \text{equation (97)} && (99)\\ &\leq C'k^{-m} && \text{Lemma 6.3} && (100)\end{aligned}$$

For n fixed and k growing ($k \gg n$), which corresponds to greater accuracy, we have that $N = \Theta(k^n)$, so from the last equation we can say there exists a constant C independent of N such that

$$E(\mathrm{W}_m^n;\mathcal{S}_{N,n}) \leq C'k^{-m} \leq CN^{-m/n} \tag{101}$$

We still have to prove that this is the (asymptotically) best possible complexity. As we have explained before, this is equivalent to prove that:

$$h_{(n+2)N}(W_m^n;c) \geq CN^{-m/n} \tag{102}$$

for some constant C. This is a direct consequence of Lemma 6.4 and Lemma 6.5. The application of both lemmas gives:

$$h_{(n+2)N}(W_m^n;c) \geq C\left(N(n+2)\right)^{-m/n} = CN^{-m/n}(n+2)^{-n/m} \tag{103}$$

This extra factor $(n+2)^{-n/m}$ is asymptotically negligible because $\lim_{n\to\infty}(n+2)^{-n/m} = 1$.

■

Corollary 6.6 (Polynomials version). *With the same hypothesis as Theorem 6.1 but restricting $f \in P_k^n$, any f can be approximated with arbitrary accuracy by shallow networks with exactly $N = (k+1)\binom{k+n-1}{k} \approx k^n$ units.*

PROOF.
This is a direct consequence of equation (97). ■

PROOF.
(Of Theorem 6.2)

$$\|f - \Sigma_f(\cdot, t)\| \leq \|f - \Sigma_f(\cdot, 0)\| + \|\Sigma_f(\cdot, 0) - \Sigma_f(\cdot, t)\| \leq \varepsilon + \|\Sigma_f(0)\|(1 - \varphi(t)) \tag{104}$$

Now we can find an upper bound for $\|\Sigma_f(0)\|$ using $f \in W_m^n$:

$$\|\Sigma_f(0)\| = \|\Sigma_f(0) - f + f\| \leq \|\Sigma_f(0) - f\| + \|f\| \leq \varepsilon + 1 \tag{105}$$

And with this result

$$\|f - \Sigma_f(\cdot, t)\| \leq \varepsilon + (\varepsilon + 1)(1 - \varphi(t)) = \varepsilon(2 - \varphi(t)) + (1 - \varphi(t)) \tag{106}$$

This is the best possible upper bound because for every n and N, there exists a function $f : \mathbb{R}^n \to \mathbb{R}$ and an approximating network $\Sigma_f \in \mathcal{S}_{n,N}$ satisfying the equality case.

To build an explicit example, we pay attention to the equality cases of each inequality used. In inequation (104) it is the triangular inequality and the fact that the accuracy is better than ε, in inequation (105) it is a combination of the same conditions before and the $\|f\|_{m,\infty} = 1$. Summarizing, we need an example where f and Σ_f are colinear, $\|f\|_{m,\infty} = 1$ and $\|\Sigma_f - f\| = \varepsilon$.

Taking everything said into consideration, given n and N, let $f \in \mathcal{S}_{n,N}$, with suitable parameters such that $\|f\|_{m,\infty} = 1$, and let $\Sigma_f = (1 - \varepsilon)f$. Then all equality cases are satisfied. ■

We would also like to have a lower bound for the error, but we have not been yet able to compute any relevant lower bound. A way of studying a lower bound would be to consider a given network of approximating error ε that we know is the best approximating network of N units. Then this network with a forgetting factor $\varphi(t)$ would also be an N-unit network, so its error by definition could not be less than ε, but we have not found any stronger version. We leave this for future work.

6.2 Deep Networks Avoid the Curse of Dimensionality

Definition 6.4 (K-ary tree, $W_m^{n,K}$ space). A CDAG is said to be a K-ary tree if each node has no more than K in-edges. We define $W_m^{n,K}$ to be a subset of W_m^n such that:

$$W_m^{n,K} \stackrel{\text{def}}{=} \left\{ f \in \mathrm{W}_m^n \ : \ \begin{array}{c} f \text{ is a } \mathcal{G}\text{-function with } \mathcal{G} \text{ a } K\text{-ary tree} \\ \text{and constituent functions } h \in \mathrm{W}_m^K \end{array} \right\} \tag{107}$$

Note that a K-ary tree is a layered graph.

The analogous theorem for deep networks is presented and proved in [39] for the case of a binary tree ($K = 2$). We state it in a more general form, but the ideas behind are essentially the same.

Theorem 6.7. *Let $\sigma \in \mathcal{C}^\infty(\mathbb{R})$ and not a polynomial. For $f \in \mathrm{W}_m^{n,K}$, the complexity of the deep networks that provide accuracy at least ϵ with the **sup** norm is*

$$N = \mathcal{O}\left((n-1)\,\epsilon^{-K/m}\right) \tag{108}$$

PROOF.
We prove this theorem by induction on d the number of hidden layers of the associated graph. The base case is equivalent to Theorem 6.1.

Consider it to be true for networks of less than d layers (induction hypothesis), we will prove the theorem for networks of exactly d layers. By Theorem 6.1 each of the constituent functions of f can be approximated up to accuracy ϵ with $\mathcal{O}\left(\epsilon^{-K/m}\right)$ units.

We wish to remark that the constituent functions of the network are Lipschitz continuous (since they are continuously differentiable in the compact set Ω). In fact, due to the norm restriction of derivatives in W_m^n and to mean value theorem, this Lipschitz constant is at most 1. We will use this fact in the proof.

If we take h to be the constituent function of the sink node, $h_1, \ldots, h_K$ the constituent functions of the layer below (the last hidden layer) and $P, P_1, \ldots, P_K$ the shallow networks approximating those mentioned constituent functions with accuracies

$$\|h - P\| \leq \frac{\epsilon}{2} \qquad \|h_i - P_i\| \leq \frac{\epsilon}{2K} \tag{109}$$

Then using Minkowskii inequality we have:

$$\|h(h_1,\ldots,h_K) - P(P_1,\ldots,P_K)\| = \|h(h_1,\ldots,h_K) - h(P_1,\ldots,P_K) + h(P_1,\ldots,P_K) - P(P_1,\ldots,P_K)\| \tag{110}$$

$$\leq \|h(h_1,\ldots,h_K) - h(P_1,\ldots,P_K)\| + \|h(P_1,\ldots,P_K) - P(P_1,\ldots,P_K)\| \tag{111}$$

The second summand, by equation (109) is less or equal than $\frac{\epsilon}{2}$. [5] The first one is bounded as follows:

$$\|h(h_1,\ldots,h_K) - h(P_1,\ldots,P_K)\| \leq \|(h_1 - P_1,\ldots,h_K - P_K)\| \quad \text{(Lipschitz)} \tag{112}$$

$$\leq \sum_{i=1}^{K} \|h_i - P_i\| \quad \text{(Triangular inequality)} \tag{113}$$

$$\leq K \cdot \frac{\epsilon}{2K} = \frac{\epsilon}{2} \quad \text{(equation (109))} \tag{114}$$

[5] It is interesting to note that here the ***sup*** norm is important. This statement will not be true in general for another $\mathscr{L}^p$ norm.

From equation (110) it directly follows that

$$\|h(h_1,\dots,h_K) - P(P_1,\dots,P_K)\| \leq \epsilon \tag{115}$$

as desired. Now by Theorem 6.1, the first approximation in equation (109) can be obtained with $\mathcal{O}\left(\left(\frac{\epsilon}{2}\right)^{-K/m}\right) = \mathcal{O}\left(\epsilon^{-K/m}\right)$ units. Since the h_i's can be considered as deep networks of $d-1$ hidden layers, each of the approximations to h_i, by induction can be obtained with $\mathcal{O}\left(\left(\frac{n}{K}-1\right)\left(\frac{\epsilon}{2K}\right)^{-K/m}\right) = \mathcal{O}\left(\left(\frac{n}{K}-1\right)\epsilon^{-K/m}\right)$.

Because there are K nodes in the final layer, the total number of units needed is indeed $\mathcal{O}\left(\epsilon^{-K/m}\right) + K\mathcal{O}\left(\left(\frac{n}{K}-1\right)\epsilon^{-K/m}\right) = \mathcal{O}\left((n-1)\epsilon^{-K/m}\right)$ ■

Both Theorem 6.1 and Theorem 6.7 are only valid for $\mathcal{C}^\infty$ activation functions. The Rectified Linear Unit (ReLU) function, which is one of the most used as an activation function, does not fall into this category. We don't believe this is a serious limitation because one can find arbitrarily close functions to it. For the case of shallow networks, very similar results have been proved for continuous (but not differentiable) activation functions considering the $\mathscr{L}^2$ norm, but they cannot be extended to deep networks using the techniques of Theorem 6.7 because the proof of Theorem 6.7 is only valid for the sup norm. We will not comment these result, the interested reader is referred to [39, Section 4].

7 Ebbinghaus Linear Forgetting Models

In this section we consider that neuron parameters are forgotten individually at some rate. We have only studied the case of two of the most used activation functions, the Rectified Linear Unit (ReLU) and the perceptron.

We will study when this approach is equivalent to our definitions in section 3 and will derive some results that give mathematically intuitive explanations about known facts in the science of forgetting.

7.1 Deterministic Model

Suppose we have a neural network that has been trained and obtained some value for the parameters (a's, $\mathbf{w}$'s and b's), as in equation (116) we will denote each variable at these "trained" values with an asterisk in the superscript (for example, a_1^*).

Now we model the forgetting of the network as these parameters being time dependent, with forgetting function $\varphi(t)$, for example: $a_1(t) = a_1^* \cdot \varphi(t)$. Typically we will think on $\varphi(t) = e^{-t/\tau}$ as the Ebbinghaus hypothesis suggests [46], but we will also analyze briefly what happens if we consider other functions instead of exponential decay.

We will assume the values of forgetting functions are between 0 and 1 as if they represent either the fraction of knowledge that is remembered or the probability that some specific knowledge is remembered.

This model is deterministic because we are not assuming that each neuron forgets at a random rate, but rather they all forget following exactly the same rate.

7.1.1 Shallow Networks: Equivalent Model

Since it is the most commonly used we will consider the activation function σ to be the ReLU function, this is, $\sigma(x) = \max(0, x)$. Since this function is linear with respect to positive number (i.e. if a is a positive number, $\sigma(ab) = a\sigma(b)$) and $\varphi(t)$ is always a positive number, it directly follows that:

$$\sigma(\langle \mathbf{x}, \mathbf{w}(t)\rangle + b(t)) = \varphi(t)\sigma(\langle \mathbf{x}, \mathbf{w}^*\rangle + b^*) \tag{116}$$

By analogy, a shallow network of the form of equation (3), becomes depending on time t as:

$$\Sigma(\mathbf{x}; t) = \sum_{k=1}^{N} a_k(t)\sigma(\langle \mathbf{x}, \mathbf{w}_k(t)\rangle + b_k(t)) \tag{117}$$

$$= \sum_{k=1}^{N} a_k^*\varphi^2(t)\sigma(\langle \mathbf{x}, \mathbf{w}_k^*\rangle + b_k^*) \tag{118}$$

$$= \varphi^2(t)\Sigma(\mathbf{x}; t = 0) \tag{119}$$

In the case of $\varphi(t) = e^{-t/\tau}$ this means that the whole networks also forgets exponentially, but with forgetting parameter $\tau/2$, which is faster than the independent neurons forget.

7.1.2 Deep Networks

Now we want to generalize to deep networks. What we will show is that if in a deep network one specific layer forgets at a given rate, say with a forgetting function $\varphi(t)$, that can be any of the ones presented before, and all biases in the layers above are tunned with the same forgetting function squared, then the whole network forgets at the same rate the specific layer is forgetting.

We propose a reasonable and simple biological mechanism for this behavior. To explain it, let us discuss the role of biases in the network. Biases can be seen as thresholds. In the ReLU case, since a neuron is activated when the logit is positive, we may assume that biases are generally negative, otherwise a neuron with null input would be activated. Therefore, the effect of multiplying by a number between 0 and 1 (only possible values of $\varphi(t)$) will actually ***increase*** the value of the bias, or equivalently, setting a lower threshold. Biologically speaking, when one layer of the network forgets with some rate, the following layers detect a lower signal than before, and so they tune their biases in a similar way the signal is lowered, to compensate the effect.

Now suppose we have a network of d layers, and the i-th layer forgets with a forgetting function $\varphi(t)$. As we have shown before in equation (119), the output of that layer will be multiplied by $\varphi^2(t)$. This is the input of the following layer, so in the $(i+1)$-th layer, the input will be multiplied by $\varphi^2(t)$. The bias will also be multiplied by $\varphi^2(t)$ by hypothesis. Using the symmetric property of the scalar product and an analogous reasoning as in equation

(119), the output of the $(i+1)$-th layer is multiplied by $\varphi^2(t)$. Repeating this reasoning with the layers above, we reach the conclusion that this $\varphi^2(t)$ affects directly to the result of the whole network.

One can prove that this behavior is additive, meaning that if another layer, say the j-th one, forgets at a given rate $\varphi'(t)$ that could be different from $\varphi(t)$, and all biases of layers $(j+1), \ldots, d$ are tuned by the way described before, then the output is multiplied by and extra $\varphi'^2(t)$. In general, if we have a different forgetting function for each layer, $\varphi_k(t)$ for $k = 1, \ldots, d$, then the whole network Σ behaves as:

$$\Sigma(\mathbf{x}; t) = \left(\prod_{k=1}^{d} \varphi_k^2(t) \right) \Sigma(\mathbf{x}; t = 0) \tag{120}$$

7.1.3 Biological Insights

7.1.3.1 More Neurons Do Not Delay Forgetting We want to note that the results obtained in this section are dependent on the number of layers, not on the number of units. This means that an increase of the number of units of a network, although it is true that improves accuracy (as seen in section 5), does not delay forgetting at all.

7.1.3.2 The Weakest Link in the Chain In the case of $\varphi_k(t)$ being exponential decay with a different time constant for each layer τ_k, equation (120) is transformed to

$$\Sigma(\mathbf{x}; t) = \left(\prod_{k=1}^{d} e^{-2t/\tau_k} \right) \Sigma(\mathbf{x}; t = 0) \tag{121}$$

$$= e^{-t\left(\sum_{i=1}^{d} \frac{2}{\tau_i}\right)} \Sigma(\mathbf{x}; t = 0) \tag{122}$$

This corresponds to an exponential forgetting in the whole network with equivalent time constant

$$\tau_{eq} = \frac{1}{\sum_{i=1}^{d} \frac{2}{\tau_i}} = \frac{1}{2d} \frac{d}{\sum_{i=1}^{d} \frac{1}{\tau_i}} \tag{123}$$

which is $\frac{1}{2d}$ times the harmonic mean of $\tau_1, \ldots, \tau_d$. This mean gives a heavier weight to small values than the common arithmetic mean, implying that the layers that forget faster (that is, with smaller τ) are the ones to determine the final forgetting behavior of the network.

7.1.3.3 Elementary Concepts Are Strongly Remembered The result in (120) may be biologically relevant to explain why it is easy for a person to forget a complex concept (for example the equation $E = mc^2$), but not to forget more basic ones (in our example, the meaning of ***equality*** and of ***multiplication***). In a deep network, complex concepts correspond to the highest layers, those closer to the output, which generally have less units than layers below. For a high layer to forget it is only needed that a few neurons forget at a given rate, while for a basic layer to forget much many neurons need to be forgetting, making it more probable for the high layer to start a forgetting process. Since each layer affects

the same way to the final result, no matter the number of neurons involved, if the network forgets some fraction of what it originally knew, we conjecture it is more probable that this forgetting comes from a high layer than from a basic one.

7.2 Random Variables Model

Now let us consider that each trainable parameter in the neuron is multiplied by a Bernoulli random variable with probability φ. This means that if a neuron is trained to have some parameters $\mathbf{w}$ and b, after training this neuron will have the same parameters with probability φ, or will have forgotten them and return 0 with probability $1-\varphi$. We will model forgetting over time varying this probability $\varphi = \varphi(t)$ with respect to time. This is a more realistic model because now each neuron forgets independently and in a random way. We can prove that in this model we get the same result as in the deterministic one in expected value.

We consider that each neuron is multiplied by a Bernoulli random variable. We want to note that this is equivalent as to multiply every trainable parameter by the same Bernoulli random variable because the ReLU activation function behaves as the identity function in all possible values of a Bernoulli random variable (which are 0 and 1). We also add to the result of each neuron a random variable, with expected value zero, to make it biologically more accurate.

We consider that all these auxiliary random variables are independent in a probabilistic sense. This hypothesis may be consistent with biological neural mechanisms related to neural plasticity if we assume that a given layer is used simultaneously and therefore forgetting mechanisms occur similarly.

On a first approach, let us consider a shallow network of N units, which results in a function on the form:

$$\Sigma(\mathbf{x}) = \sum_{k=1}^{N} a_k\big(\sigma(\langle \mathbf{x}, \mathbf{w}_k\rangle + b_k)\beta_k + \delta_k\big) \tag{124}$$

where β_k are the Bernoulli random variables and δ_k are the ones with $\mathbb{E}(\delta_k) = 0$. We are interested in the expected value of Σ, which we denote as $\mathbb{E}(\Sigma)$. Using linearity of the expected value on (125) and the expected values of β_k's and δ_k's on (126), it follows:

$$\mathbb{E}(\Sigma(\mathbf{x};t)) = \sum_{k=1}^{N} a_k\left(\sigma(\langle \mathbf{x}, \mathbf{w}_k\rangle + b_k)\mathbb{E}(\beta_k) + \mathbb{E}(\delta_k)\right) \tag{125}$$

$$= \varphi(t)\sum_{k=1}^{N} a_k\sigma(\langle \mathbf{x}, \mathbf{w}_k\rangle + b_k) \tag{126}$$

Since at $\varphi(t=0) = 1$, which corresponds to the original trained network, we can say that $\Sigma(\mathbf{x}; t=0) = \sum_{k=1}^{N} a_k\sigma(\langle \mathbf{x}, \mathbf{w}_k\rangle + b_k)$, and thus the previous result can be synthesized as:

$$\mathbb{E}(\Sigma(\mathbf{x};t)) = \varphi(t)\mathbb{E}\Big(\Sigma(\mathbf{x};t=0)\Big) \tag{127}$$

In the generalization to deep networks, we will make an analogous hypothesis as in the deterministic case for the biases. We consider that, whenever the i-th layer is forgotten with a

forgetting function $\varphi(t)$, all biases in the layers above, $(i+1)$-th until the output layer, forget their biases in a similar way. Mathematically, they are multiplied by a Bernoulli random variable of probability $\varphi(t)$. We will not consider the random variables δ_k's for the deep network case. Also, we need to assume that for layers $i, \dots, d$, the parameters a_k's and $\mathbf{w}_k$'s are all non-negative.

The result we get is that, if a layer, say the i-th layer is forgotten with probability function $\varphi(t)$, and biases fulfill the previous hypothesis, then the expected value of the whole network fulfills the following inequality:

$$\mathbb{E}(\Sigma(\mathbf{x};t)) \geq \varphi(t)\Sigma(\mathbf{x};t=0) \tag{128}$$

PROOF.
We prove this by induction on the number of layers of the network d. The base case corresponds to $d=1$, and has been already proved in (127).

Consider a network of d layers in which the i-th layer has been forgotten (for some $i<d$). If $i=d-1$, the output of the network is:

$$\Sigma^d(\mathbf{x}) = \sum_{k=1}^{N^{(d-1)}} a_k^{(d-1)} \left(\sigma\left(\left\langle \Sigma^{d-1}(\mathbf{x}), \mathbf{w}_k^{(d-1)} \right\rangle + b_k^{(d-1)}\right)\beta_k^{(d-1)}\right) \tag{129}$$

Where the $\beta_k^{(d-1)}$'s are associated with the forgetting in the $(d-1)$-th layer, and $\Sigma^{(d-1)}(\mathbf{x})$ is not a random variable. By analogy with the shallow network case, $\mathbb{E}\Big(\Sigma^d(\mathbf{x};t)\Big) = \varphi(t)\Sigma^d(\mathbf{x};t=0)$, which is in fact an equality case.

If $i<d-1$, the output of such network will be of the form (the superscripts account for the layer):

$$\Sigma^d(\mathbf{x}) = \sum_{k=1}^{N^{(d-1)}} a_k^{(d-1)} \sigma\left(\left\langle \Sigma^{d-1}(\mathbf{x}), \mathbf{w}_k^{(d-1)} \right\rangle + b_k^{(d-1)}\beta_k^{(d-1)}\right) \tag{130}$$

In this equation, there are two types of random variables: $\Sigma^{d-1}(\mathbf{x})$ and the β_k^d's. The first comes from the randomness in the layers below, and is only random if $i<d-1$, while the β_k^d's are just Bernoulli random variables of some probability φ.
Applying linearity of expected value it follows that:

$$\mathbb{E}\Big(\Sigma^d(\mathbf{x})\Big) = \sum_{k=1}^{N^{(d-1)}} a_k^{(d-1)} \mathbb{E}\Big(\sigma\left(\left\langle \Sigma^{d-1}(\mathbf{x}), \mathbf{w}_k^{(d-1)} \right\rangle + b_k^{(d-1)}\beta_k^{(d-1)}\right)\Big) \tag{131}$$

Since $i<d-1$ and $\Sigma^{(d-1)}(\mathbf{x})$ is the output of a network with $d-1$ layers, the induction hypothesis tells us that $\mathbb{E}(\Sigma^{(d-1)}(\mathbf{x};t)) \geq \varphi(t)\Sigma^{(d-1)}(\mathbf{x};t=0)$. In order to apply the induction hypothesis, we want that $\mathbb{E}\Big(\sigma(\langle \Sigma^{(d-1)}, \mathbf{w}_k^{(d-1)}\rangle + b_k^{(d-1)}\beta_k^{(d-1)})\Big) \geq \sigma(\mathbb{E}(\langle \Sigma^{(d-1)}, \mathbf{w}_k^{(d-1)}\rangle + b_k^{(d-1)}\beta_k^{(d-1)})\Big)$, which is actually true as we will see.

As a notation convention, let us define the random variable Y as $Y = \langle \Sigma^{(d-1)}, \mathbf{w}_k^{(d-1)}\rangle + b_k^{(d-1)}\beta_k^{(d-1)}$. Since the randomness comes from the effect of a finite number of Bernoulli

random variables, given the input $\mathbf{x}$ the variable Y can only take a finite number of values. Let S be the (finite) set of all possible values for Y, let S^+ be the set of all non-negative members of S and S^- the set of all negative values of S. Using the definition of expected value in (132), the property that $\sigma(y) = 0$ when $y \leq 0$ in (133), the linearity of σ for non-negative values and the fact that the second sum is zero in (134), the property that $\sigma(a) + \sigma(b) \geq \sigma(a + b)$ in (135) results in:

$$\mathbb{E}\Big(\sigma(Y)\Big) = \sum_{y \in S} Pr(Y = y)\sigma(y) \tag{132}$$

$$= \sum_{y \in S^+} Pr(Y = y)\sigma(y) \tag{133}$$

$$= \sigma\left(\sum_{y \in S^+} Pr(Y = y) \cdot y\right) + \sigma\left(\sum_{y \in S^-} Pr(Y = y) \cdot y\right) \tag{134}$$

$$\geq \sigma\Big(\sum_{y \in S^+ \cup S^-} Pr(Y = y)y\Big) \tag{135}$$

$$= \sigma\Big(\mathbb{E}(Y)\Big) \tag{136}$$

If we apply this result to (131) and the induction hypothesis as said the proof is done. ■

This tells us that under the stated hypothesis, forgetting does not happen faster than in the deterministic case, but it could be slower, since we have not found an upper bound for $\mathbb{E}(\Sigma(\mathbf{x}; t))$ in terms of $\mathbb{E}(\Sigma(\mathbf{x}; t = 0))$.

8 Conclusion and Future Research

There is not conclusive evidence that all animal memory can be modeled as stored in deep neural networks and there still exist apparently-unsurmountable differences between deep neural networks and human memory (most notably, humans learn with few examples), however, we have proven forgetting deep neural networks can be part of the puzzle given the remarkable similarities between how they forget and the way animals do it – in fact, in a parallel review effort [46] we have found no single piece of evidence reviewed contradicting a conjecture that our model is part of how the human brain forgets.

The similarities between deep forgetting neural networks and human memory forgetting are remarkable as summarized here:

(1) It's not instantaneous (see subsection 4.1)

(2) It's unavoidable (see subsection 4.2)

(3) Reinforcement delays forgetting (see subsection 2.5)

(4) More neurons do not increase retention (see paragraph 7.1.3.1)

(5) Forgetting of "features" propagates to "concepts" (see paragraph 7.1.3.2)

(6) "Concepts" are forgotten faster than "features" see (paragraph 7.1.3.3). Reference amnesia example – i.e. things are different.

(7) "High frequency" memories are lost faster than "low frequency" ones (see subsection 5.4)

(8) Forgetting is isolated to areas without reinforcement (see subsection 2.5)

(9) Forgetting occurs at Ebingghaus speed.

(10) Some "genetic" knowledge may be "hard-wired" and never be forgotten.

This is certainly a partial list and is presented here just to illustrate the many potential avenues of possible future research.

Our ultimate goal is to find a model of the learning and forgetting dynamics in the animal brain. We postulate what we call a *dynamic forgetting network*, which consists on a triple $(\mathcal{N}, \varphi, \mathcal{A})$, where $\mathcal{N}$ is a regular neural network, φ is a set of forgetting functions and $\mathcal{A}$ is a learning algorithm (given an input $(\boldsymbol{x}, f(\boldsymbol{x}))$ modifies the parameters of $\mathcal{N}$). So far we have shown that our model is somewhat consistent with 10 important characteristics of forgetting in the human brain, known experimentally, and have not found any such characteristic that is inconsistent with our model.

In the rest of this conclusion we present five areas of future research we believe will be essential towards the goal above.

8.1 Optimizing Choice of Learning Algorithms in Forgetting Networks

We have focused on forgetting mechanisms (not learning algorithms). There is a huge variety of learning algorithms, and it is unclear which one of them could best be suited to a forgetting model of the animal brain. We think modern learning techniques used in deep learning algorithms can be introduced in our model, and hope this contributes to our understanding of the human brain.

8.2 Modeling Forgetting beyond Weight Loss

We have conducted extensive experiments with real neural network implementations to demonstrate the behavior of neural networks when weights are modified, and the results (to be reported more elsewhere ([45]) are consistent with the notion that decreases in weight translate into memory loss – however, our experiments also show that other possible mechanisms at the neural level may be playing a role.

For example, in our empirical experiments, we have tested the impact of lowering the bias as shown in Figure 10 [6] or changing the percentage of neurons in a layer set to zero

[6] The experiments where done on the MNIST dataset, where the input consists of 60000 one- channel 28x28 images, and the output of 10 values of probability, for each digit (ranging from 0 to 9). The model that has been used is, from input to output:

- Convolutional layer of 3x3 kernel size and ReLU activation, with 32 filters.

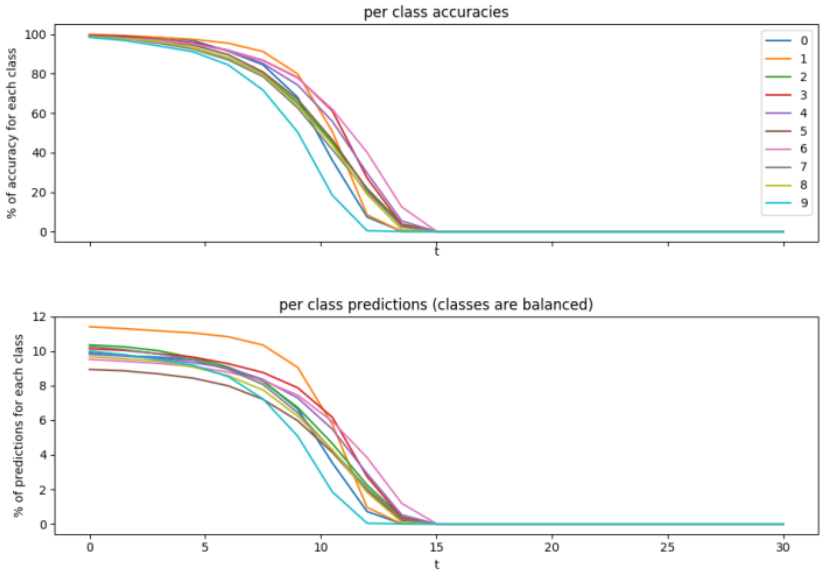

Figure 10: For some activation functions modifying the bias may also be a way of modeling forgetting. This experiment was conducted on...

suggesting that perhaps more than one neural mechanism may be playing a role in forgetting – further mathematical modeling is required to understand the impact on forgetting of the various components of a forgetting deep neural network architecture.

- Convolutional layer of 3x3 kernel size and ReLU activation, with 64 filters.
- MaxPooling2D layer of 2x2 pool size, with a Dropout of 0.25 probability.
- Fully Connected layer with 128 units, tanh activation, and a Dropout of 0.5 probability.
- Fully Connected layer with 10 units and softmax activation, that produces the outputs.

This model, trained for 12 epochs, yields an accuracy of 99.07% on the test set, of 10000 images.

Keras has been used as the framework to develop the experiments. Because of its conventions when describing the network, the following layers where used:

- First convolutional layer is #1.
- Second convolutional layer is #2.
- First FC layer is #6.
- Second FC layer BEFORE the softmax function is #8.
- Second FC layer AFTER the softmax function is #9.

8.3 Stochastic Modeling of Forgetting

We also want to work towards a stochastic version of the model. In that direction, we have explored the following basic model: consider that the activation function σ is a perceptron with some threshold T

$$\sigma(x) = \begin{cases} 1 & \text{if } x \geq T \\ 0 & \text{if } x < T \end{cases} \tag{137}$$

Consider a network with input $\boldsymbol{x} \in \mathbb{R}^n$ on the form of (3). Now consider that, given an input $\boldsymbol{x}$, $\langle \boldsymbol{x}, \boldsymbol{w} \rangle$ is a random variable following a binomial distribution $\text{Bin}(n, p)$, where $p = \varphi(t)$ is the Ebbinghauss curve. We consider σ to have threshold $n/2$ and all biases set to zero. We will call each neuron in this setup a ***synaptic vesicle*** $\text{sv}(n, p, n/2)$. In this context a shallow network becomes:

$$\Sigma(\boldsymbol{x}) = \sum_{k=1}^{N} \sigma(\langle \boldsymbol{x}, \boldsymbol{w} \rangle) \sim \text{Bin}(n, \tilde{p}) \tag{138}$$

where $\tilde{p}$ is the probability that each neuron has to activate. Each neuron activates when the value of the logit is greater or equal to $n/2$, thus

$$\tilde{p} = \sum_{k=\lceil n/2 \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k} \tag{139}$$

Although we have not found a more compact formula for $\tilde{p}$, we have found experimentally that if $p > 1/2$, then $\tilde{p} > p$, and if $p < 1/2$, then $\tilde{p} < p$.

The generalization to deep networks is done by considering that the logits of the layer l follow a binomial distribution $\text{Bin}(n, p_l)$, where $p_l = \tilde{p_{l-1}}$, which corresponds to the distribution of the outputs of the previous layer. With such generalization, given $p_1 = p$ some probability of forgetting, it "radicalizes" as it goes through the network: if $p_{initial} > 1/2$, p_{final} is close to 1, whereas if $p_{initial} < 1/2$, p_{final} is close to 0, meaning that there is a point ($p = 1/2$) at which there is a strong change in behavior, corresponding to the time the network effectively forgets.

In future research we want to model more complicated networks via addition of many synaptic vesicles. We have proven the following rule: If you add N synaptic vesicles $\text{sv}(n, p, n/2)$ and then the result is passed through a perceptron of threshold 1, then the result follows a Bernoulli distribution $B(q)$ where $q = 1 - (1 - \tilde{p})^N$, where $\tilde{p}$ is the one in equation (139).

8.4 Explaining Practice Scheduling and Forgetting learning

Figure 11 illustrates various learning practices that have an impact on memory saliency. Future research could try to pair observed retention in humans with alternative neural forgetting mechanisms to help elucidate the apropppriate explanations for each case, including the role of various forgetting functions and learning strategies.

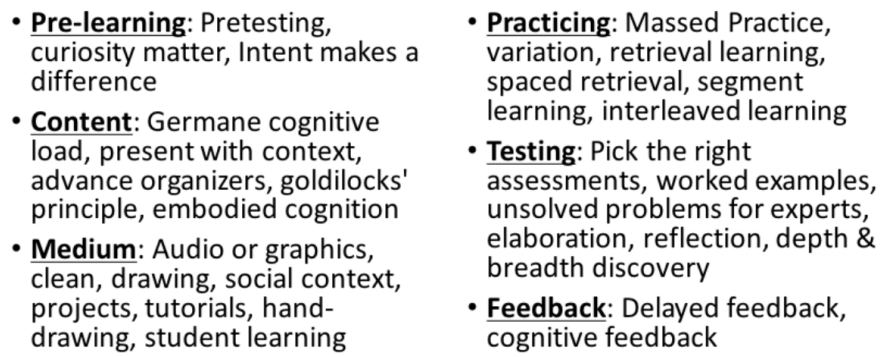

Figure 11: This figure illustrates various learning practices that have been reviewed, without finding any evidence of a long-term retention experiment that links performance with a given practice. The methodologies have been organized using a novel taxonomy presented in [46].

The forgetting curve may be shaped by factors that have not been reviewed in the literature, perhaps not even addressed or ever suggested including a more complex set of values (e.g. affection and social interaction [36]), memory flaws [43], type of memory [41, 48]. Wixted [49] suggests that perhaps there is a different behavior between memories that have not consolidated and those that have. Another related aspect that has not been addressed is when is memory "refreshed" or "practiced", in other words, when should we re-set the clock to zero. If we review derivatives, are we also implicitly reviewing integrals? If we review how the Pythagoras theorem can be derived, are we also reviewing how it can be applied? Can forgetting be accelerated? All of these possibilities may be incorporated in forgetting deep learning networks to provide further insights into human forgetting and perhaps even more connections between the two.

8.5 Explaining The Baby Forgetting Conjecture: "Babies Don't Really Learn and Mostly Forget"

There is some controversy as to what is learned versus what is pre-wired in the animal brain, and more specifically in humans, especially as it relates to the first development phases of a baby's life. There are many experiments on several aspects of primate baby "learning", with no conclusive answers to basic questions such as why a baby "learns so quickly" and more research is continually being proposed. We, instead, suggest what is really needed is a different theory.

Our suggested theory, which we call the "baby forgetting hypothesis" is that the key mechanism behind many so-called "learning" early phenomena is, in fact, a forgetting one, one which selects among many potential feature sets based on a model that corresponds, to a certain extent, to a drawing similar to that of Figure 7, Figure 8 and Figure 9. We

have shown that once the basis of a polynomial has been constructed, new polynomials recombining the basis can be constructed just with one neuron each (i.e. in a computationally very efficient way). This would mean that the baby brain may start at birth with, say, having for a given sensory region one thousand polynomials of order 5 and then, over the first phases of life, prune the ones that do not appear in the perceived feature set. A similar mechanism, possibly with different default σ's, "base" polynomials, and forgetting schedule may be at stake in different parts of the brain so that evolution may simply be selecting the number of neurons and which of these few parameters to optimize in different parts of the brain. There is evidence supporting this embryonic hypothesis.

Hubel and Wisel's cat experiments [21] are consistent with a model in which orientation is ready to be detected when seen – but only until week twelve. Cats don't perceive vertical lines if they haven't seen them but they can quickly learn to see them (in minutes) if they are exposed to vertical lines, but only up until that week. In other words, after that time the cat will never be able to develop that skill no matter how many examples it sees. A similar phenomenon may happen with stereo or human face recognition. "Feature" selectivity may in fact also be the same mechanism for higher level concepts. Recent evidence on primates has shown they ignore faces unless they have been exposed to them [4, 40]. The limited preference for faces over other stimuli in babies, only developed in non-pre-term babies [35], may well be an artifact of the feature set babies are born with instead of a "two-process" theory as suggested elsewhere [34]. The concept of minimal images [47] may also be based on aggregating basic features into a composed representation. The behavior of neural nets and humans has been shown [47] to be different with minimal images, but if there was a feature layer, the behavior could be the same. For this to be true, lower layers would calculate the features that compose a minimal image and be useless for more degraded images. This could result in a model where minimal images in forgetting neural networks could reproduce the behavior of humans. This would also greatly improve the modeling power of such networks since the catastrophic forgetting exhibited by neural networks raises a lot of questions [17, 23].

Vision is not the only domain where there may be support for the "baby forgetting hypothesis". In speech, it has been shown that, like in the case of cat's or minimal images above, humans first learn the accent and then the language [26]. Babies have also shown preference for the sounds of the mother [[12] which may be the result of "forgetting" the other features (instead of the perhaps more intuitive notion that they somehow "learn" the mother's voice). The rapid subliminal reaction to human expressions [13] may also be the result of inter-modal connections at the early layers of perception based on feature selectivity across parts of the brain.

We don't mean to imply by our conjecture that the brain does not have any "meta learning" hability starting at birth, but instead that perhaps many of the evidence shown experimentally can be explain by an underlying selection or forgetting mechanism that has nothing to do with backpropagation and that may work on whatever experiments are available. We feel at this point we know to little to make any claims about where is the frontier between what is selective forgetting versus what is learned via a different mechanism in the early stages of life. In fact, it is possible that even the developmnet of grit [14], perhaps one of the most relevant open problems in learning science, is also based on some form of forgetting

given that babies learn the value of effort early on [31] and there is evidence that they show the grit they see in their parents already when they are 15 months [28] (the analogy with Hubel and Weisel cat experiments being that the brain exhibits a response based on very fewo examples of the behavior it perceives and "forgets" the one it does not).

The forgetting theorems proven are consistent with a model where "learning" really starts by "forgetting" in a well-planned, region-specific and structured way pre-programmed at birth – this is only a "directional" conjecture and more research is required to validate it empirically, by implementing forgetting neural networks and, mathematically, by solving basic open research problem such as "what may be the theoretically optimal basis upon which to start a forgetting process", and subsequently, "what are the associated techniques that can prune available options" (which may be based on gradient descent, or who knows, in other biologically more plausible mechanisms such as the ones described in [15, 16]). More research is also necessary to understand whether neural selectivity and consistency across individuals is an artifact of basic feature selectivity or, instead, a more pre-coded born-with skill [8].

Ultimately, no model of human learning will be complete if it does not take into account the role of forgetting in the baby's brain development.

References

[1] A. DeVore, R., Howard, R., and Micchelli, C. Optimal nonlinear approximation. *Manuscripta Mathematica 63* (12 1989), 469–478.

[2] Adams, R., and Fournier, J. *Sobolev Spaces*. Pure and Applied Mathematics. Elsevier Science, 2003.

[3] Anselmi, F., Rosasco, L., Tan, C., and Poggio, T. Memo 35: Deep convolutional networks are hierarchical kernel machines. *Center for Brains, Minds and Machines* (2015).

[4] Arcaro, M. J., et al. Seeing faces is necessary for face-domain formation. *Nature Neuroscience 20*, 10 (2017), 1404–1412.

[5] Arnol'd, V. I. On the representation of continuous functions of three variables as superpositions of continuous functions of two variables (in Russian). *Dokl. Akad. Nauk SSSR 114*, 4 (1957), 679–681.

[6] Bagby, T., Bos, L., and Levenberg, N. Multivariate simultaneous approximation. *Constructive Approximation 18*, 4 (Dec 2002), 569–577.

[7] Baire, R. Sur les fonctions de variables réelles. *Annali di Matematica Pura ed Applicata (1898-1922) 3*, 1 (Dec 1899), 1–123.

[8] Barraclough, N. E., and Perrett, D. I. From single cells to social perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences 366*, 1571 (2011), 1739–1752.

[9] Buchberger, B. Phd thesis 1965: An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal. *Journal of Symbolic Computation 41*, 3 (2006), 475 – 511. Logic, Mathematics and Computer Science: Interactions in honor of Bruno Buchberger (60th birthday).

[10] Buhmann, M. Radial basis function. *Scholarpedia 5*, 5 (2010), 9837. revision #137035.

[11] Burden, R., and Faires, J. *Numerical Analysis*. Cengage Learning, 2010.

[12] DeCasper, A. J., and Fifer, W. P. Of human bonding: Newborns prefer their mothers' voices. *Science 208*, 4448 (2016), 1174–1176.

[13] Dimberg, U., Thunberg, M., and Elmehed, K. Unconscious facial reactions to emotional facial expressions. *Psychological science 11*, 1 (2000), 86–89.

[14] Duckworth, A. L., Peterson, C., Matthews, M. D., and Kelly, D. R. Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology 92*, 6 (2007), 1087.

[15] Edelman, G. M. Neural Darwinism: selection and reentrant signaling in higher brain function. *Neuron 10*, 2 (1993), 115–125.

[16] Edelman, G. M., and Mountcastle, V. B. *The Mindful Brain: Cortical Organization and the Group-Selective Theory of Higher Brain Function*. MIT Press, 1982.

[17] French, R. M. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences 3*, 4 (1999), 128–135.

[18] Girosi, F., and Poggio, T. Representation Properties of Networks: Kolmogorov's Theorem Is Irrelevant. *Neural Computation 1*, 4 (Dec. 1989), 465–469.

[19] Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks. *ArXiv e-prints* (Dec. 2013).

[20] Hörmander, L. *The analysis of linear partial differential operators: Distribution theory and Fourier analysis*. Springer Study Edition. Springer-Verlag, 1990.

[21] Hubel, D. H., and Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology 160*, 1 (1962), 106–154.

[22] Jackson, D. On approximation by trigonometric sums and polynomials. *Transactions of the American Mathematical Society 13*, 4 (1912), 491–515.

[23] Kirkpatrick, J., et al. Overcoming catastrophic forgetting in neural networks. *CoRR abs/1612.00796* (2016).

[24] Kolmogorov, A. N. On the representation of continuous functions of several variables as superpositions of continuous functions of a smaller number of variables (in Russian). *Dokl. Akad. Nauk SSSR 108*, 2 (1956), 179–182.

[25] Kůrková, V. Kolmogorov's Theorem is Relevant. *Neural Comput. 3*, 4 (Dec. 1991), 617–622.

[26] Kuhl, P. K. Learning representation in speech and language. *Current opinion in neurobiology 4*, 6 (1994), 812–822.

[27] Lang, S. *Real and Functional Analysis*, 3 ed. Graduate Texts in Mathematics 142. Springer-Verlag New York, 1993.

[28] Leonard, J. A., Lee, Y., and Schulz, L. E. Infants make more attempts to achieve a goal when they see adults persist. *Science 357*, 6357 (2017), 1290–1294.

[29] Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks 6*, 6 (1993), 861 – 867.

[30] Lin, V., and Pinkus, A. Fundamentality of ridge functions. *Journal of Approximation Theory 75*, 3 (1993), 295 – 311.

[31] Liu, S., Ullman, T. D., Tenenbaum, J. B., and Spelke, E. S. Ten-month-old infants infer the value of goals from the costs of actions. *Science 358*, 6366 (2017), 1038–1041.

[32] Marr, D. *Vision: A computational investigation into the human representation and processing of visual information*. WH San Francisco: Freeman and Company, 1982.

[33] Mhaskar, H. N. Neural Networks for Optimal Approximation of Smooth and Analytic Functions. *Neural Computation 8* (1996), 164–177.

[34] Morton, J., and Johnson, M. H. CONSPEC and CONLERN: a two-process theory of infant face recognition. *Psychological review 98*, 2 (1991), 164.

[35] Pereira, S. A., et al. A comparison between preterm and full-term infants' preference for faces. *Jornal de pediatria 93*, 1 (2017), 35–39.

[36] Picard, R., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D., and Strohecker, C. Affective learning—a manifesto. *BT technology journal 22*, 4 (2004), 253–269.

[37] Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta Numerica 8* (1999), 143–195.

[38] Poggio, T., and Liao, Q. Memo 66: Theory ii: Landscape of the empirical risk in deep learning. *Center for Brains, Minds and Machines* (2017).

[39] Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., and Liao, Q. Memo 58: Why and when can deep - but not shallow - networks avoid the curse of dimensionality: a review. *Center for Brains, Minds and Machines* (2016).

[40] Sackett, G. P. Monkeys reared in isolation with pictures as visual input: Evidence for an innate releasing mechanism. *Science 154*, 3755 (1966), 1468–1473.

[41] Sadeh, T., Ozubko, J., Winocur, G., and Moscovitch, M. Forgetting Patterns Differentiate Between Two Forms of Memory Representation. *Psychological science 27*, 6 (2016), 810–820.

[42] Salsa, S. *Partial Differential Equations in Action: From Modelling to Theory*. Universitext. Springer Milan, 2008.

[43] Schacter, D. *The seven sins of memory: How the mind forgets and remembers*. Houghton Mifflin Harcourt, 2002.

[44] Scholkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing 45*, 11 (1997), 2758–2765.

[45] Subirana, and Rotger. forthcoming (2018).

[46] Subirana, B., Bagiati, A., and Sarma, S. Memo 68: On the forgetting of college academics: at "ebbinghaus speed"? *Center for Brains, Minds and Machines* (2017).

[47] Ullman, S., et al. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences 113*, 10 (2016), 2744–2749.

[48] Weimer-Stuckmann, G. *Second Language Vocabulary Acquisition: Spacing and Frequency of Rehearsals*. Master's dissertation, University of Victoria, 2009.

[49] Wixted, J. A theory about why we forget what we once knew. *Current Directions in Psychological Science 14*, 1 (2005), 6–9.

[50] Zhang, C., Liao, Q., Rakhlin, A., Sridharan, K., Miranda, B., Golowich, N., and Poggio, T. Memo 67: Theory of deep learning iii: Generalization properties of sgd. *Center for Brains, Minds and Machines* (2017).

Appendices

A Proofs

A.1 Proof of Lemma 5.3

PROOF.

$\boxed{i \implies ii}$

Let m be the degree of σ, then its $(m+1)$-th derivative vanishes, so $\forall x \in \mathbb{R}$, $\sigma^{(m+1}(x) = 0$.

$\boxed{ii \implies i}$

Let us define $E_n \stackrel{\text{def}}{=} \{x \in \mathbb{R} \,:\, f^{(n}(x) = 0\}$. This sets are closed due to the continuity of f and $\cup_{n\in\mathbb{N}} E_n = \mathbb{R}$ by hypothesis. We recall Baire's category theorem ([7]) which in our case states that for every numerable collection of dense open sets $\{U_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}$, their intersection is dense. In our case, let U_n be the complementary set of E_n: $U_n \stackrel{\text{def}}{=} E_n^c$. If all U_n were dense, by Baire's theorem, its intersection would be dense. But taking into account De Morgan laws

$$\bigcap_{n\in\mathbb{N}} U_n = \bigcap_{n\in\mathbb{N}} E_n^c = \left(\bigcup_{n\in\mathbb{N}} E_n\right)^c = \mathbb{R}^c = \emptyset \quad \text{(which is not dense in } \mathbb{R}) \tag{140}$$

As a consequence, there is one set U_n not dense, equivalently, one set E_n with non-empty interior, therefore it contains an interval $I \subseteq E_n$. So we have an interval I where $f^{(n}(x) = 0 \quad \forall x \in I$, meaning that f is a polynomial of degree at most n in I.

Now let Λ be a set of indexes, and $\{I_\lambda\}_{\lambda\in\Lambda}$ be the set of all maximal open intervals I_λ such that f is a polynomial in I_λ. We have already seen that there exists at least one such interval. We also observe that these intervals are mutually disjoint because they are maximal (if two I_λ and I_μ satisfy $I_\lambda \cap I_\mu \neq \emptyset$, then $I_\lambda \cup I_\mu$ is a larger interval).

Now we define

$$H \stackrel{\text{def}}{=} \mathbb{R} \setminus \bigcup_{\lambda\in\Lambda} I_\lambda \tag{141}$$

H has empty interior. If it did not, it would contain an interval $J \subseteq H$ and applying the same argument with Baire's theorem as previously to J instead of $\mathbb{R}$ we would find an interval $J' \subseteq J$ in which f is a polynomial, which generates a contradiction.

Now we prove that H has no isolated points. Suppose $x \in H$ is an isolated point. For this to happen, there would be two intervals $I_1, I_2 \in \{I_\lambda\}_{\lambda\in\Lambda}$ such that x is the right endpoint of I_1 and the left endpoint of I_2. There would then be also an integer n such that the n-th derivative vanishes in $I_1 \cup I_2$, and by continuity of $f^{(n}$ it would also vanish in x, and thus $I_1 \cup \{x\} \cup I_2$ is a larger interval in which f is a polynomial.

H is a closed subspace of $\mathbb{R}$, so if it is not empty, using Baire's theorem again [7], there

[7] $E_n \cap H$ are closed subsets of H and $\bigcup_{n\in\mathbb{N}}(E_n \cap H) = H$, then there must be some of the E_n having non-empty interior in H.

exists and interval J such that $J \cap H \neq \emptyset$ and for some n

$$f^{(n}(x) = 0 \qquad \forall x \in J \cap H \tag{142}$$

Since H has non isolated points, any $x \in J \cap H$ is an accumulation point of $J \cap H$, so using only points in $J \cap H$ we can calculate

$$\lim_{h \to 0} \frac{f^{(n}(x+h) - f^{(n}(x)}{h} \tag{143}$$

This limit exists because f is infinitely differentiable and by construction it vanishes in $J \cap H$

$$f^{(n+1}(x) = 0 \qquad \forall x \in J \cap H \tag{144}$$

and repeating this argument,

$$f^{(m}(x) = 0 \qquad \forall x \in J \cap H \quad \forall m \geq n \tag{145}$$

Now we claim that there exists an interval $I \in \{I_\lambda\}_{\lambda \in \Lambda}$ contained in J. If there were no such interval, it would mean that the set $J \cap H$ contains an interval. But then H would contain an interval in which f is a polynomial, that should have been included in $\{I_\lambda\}_{\lambda \in \Lambda}$.

Let $I \subseteq J$ be such interval. Since f is a polynomial in I, there exists m such that $f^{(m}$ vanishes in I. Suppose $m > n$. Since the endpoints of I are in $J \cap H$ (therefore $f^{(m}$ vanishes at the mentioned endpoints) and $f^{(m} = 0$ in I, it is deduced that $f^{(m-1}$ vanishes in I. Applying induction we deduce that $f^{(n}$ vanishes in I.

Choose a point $x \in J \cap H$. We have seen that $J \cap H$ cannot contain an interval, so there must exist two intervals $I_1, I_2 \in \{I_\lambda\}_{\lambda \in \Lambda}$ such that x is the left endpoint of I_2 and the right endpoint of I_1. Since $f^{(n}$ is continuous and is zero at $I_1 \cup I_2$, it must be zero also in x, which is a contradiction because then f is a polynomial in $I_1 \cup \{x\} I_2$ which is a larger interval. This contradiction comes from assuming $H \neq \emptyset$, and f is then a polynomial in the only maximal interval $\mathbb{R}$. ■

A.2 Details of proof of Theorem 5.7

Suppose we have a sequence of polynomials of degree at most k: $\{P_i(\cdot)\}_{i \in \mathbb{N}}$, $P_i(x) = \sum_{j=1}^{k} a_j^i x^j$ such that

$$P_i(\cdot) \xrightarrow[i \to \infty]{} f$$

We consider two possibilities:

i $\forall j$, $\exists a^j \in \mathbb{R}$ such that $a_j^i \xrightarrow[i \to \infty]{} a_j$. In this case we will prove that $f(x) = a_0 + a_1 x + \cdots + a_k x^k$.

ii $\exists j_0$ such that $\{a_{j_0}^i\}_{i \in \mathbb{N}}$ has no limit. We will prove this case is in contradiction with $\{P_i\}_{i \in \mathbb{N}}$ being convergent.

In the first case, let us define $M = \max_{j=0,\dots,k} \|x^j\|$. Given $\varepsilon > 0$, we can choose i_0 such that

$$|a_j^i - a_j| < \frac{\varepsilon}{(k+1)M} \quad \forall i \geq i_0 \, \forall j = 0, \dots, k \tag{146}$$

If we define $g(x) = a_0 + a_1 x + \cdots + a_k x^k$, if follows $\forall i > i_0$:

$$\begin{aligned}
\|g - P_i\| &= \|(a_0 - a_0^i) + \cdots + (a_k - a_k^i)x^k\| && && (147)\\
&\leq |a_0 - a_0^i| + \cdots + |a_k - a_k^i|\|x^k\| && \text{Triangular inequality} && (148)\\
&\leq (|a_0 - a_0^i| + \cdots + |a_k - a_k^i|)M && \text{Definition of } M && (149)\\
&\leq \left(\frac{\varepsilon}{(k+1)M} + \cdots + \frac{\varepsilon}{(k+1)M}\right) M = \varepsilon && \text{equation (146)} && (150)
\end{aligned}$$

In conclusion, $P_i \xrightarrow[i\to\infty]{} g$, since the limit is unique, $f = g$.

In case (ii), since $P_i \xrightarrow[i\to\infty]{} f$ implies $P_i(x) \xrightarrow[i\to\infty]{} f(x)$ almost for any x, we can choose $k+1$ different numbers $y_0, \dots, y_k$ such that the convergence is pointwise, i.e.:

$$\begin{cases} a_0^i + a_1^i y_0 + \cdots + a_k^i y_0^k & \xrightarrow[i\to\infty]{} \quad f(y_0) \\ \vdots & \quad\vdots \\ a_0^i + a_1^i y_k + \cdots + a_k^i y_k^k & \xrightarrow[i\to\infty]{} \quad f(y_k) \end{cases} \tag{151}$$

Using properties of limits, we can make a linear combination of these limits to get:

$$a_0^i \left(\sum_{j=0}^k \gamma_j\right) + a_1^i \left(\sum_{j=0}^k \gamma_j y_j\right) + \dots a_k^i \left(\sum_{j=0}^k \gamma_j y_j^k\right) \xrightarrow[i\to\infty]{} \sum_{j=0}^k \gamma_j f(y_j) \tag{152}$$

where γ_j are coefficients of the linear combination. If we can choose these coefficients such that

$$\sum_{j=0}^k \gamma_j y_j^l = \begin{cases} 1 & \text{if } l = j_0 \\ 0 & \text{otherwise} \end{cases} \tag{153}$$

equation (152) becomes $a_{j_0}^i \xrightarrow[i\to\infty]{} \sum_{j=0}^k \gamma_j f(y_j)$ which is a contradiction.
It is indeed possible to choose $\{\gamma_j\}_{j=0:k}$ as in equation (153), because they are the solution of the system of equations

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ y_0 & y_1 & \dots & y_k \\ \vdots & & & \vdots \\ y_0^k & y_1^k & \dots & y_k^k \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow j_0\text{-th position} \tag{154}$$

The matrix of this system of equations is a Vandermonde matrix, therefore it has a unique solution if and only if all y_j are different.

We have proved that a sequence of polynomials of degree at most k can only have a polynomial of degree at most k as a limit.

A.3 Proof of Lemma 6.3

Lemma 6.3 is a classical result whose proof is based on Jackson's Theorem (see Theorem A.1). It involves some concepts and previous results of approximation theory that will explained in this section.

A.3.1 Derivation from Jackson's Theorem

We begin by presenting the concepts of ***modulus of smoothness*** and ***best approximation*** by a polynomial.

Definition A.1 (Modulus of smoothness). Given $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$, with partial derivatives up to order m, the ***modulus of smoothness of*** f ***of order*** m is a function $\omega_{f,m} : [0, \infty) \to [0, \infty)$ defined by

$$\omega_{f,m}(\delta) \stackrel{\text{def}}{=} \sup_{|\boldsymbol{\gamma}|=m} \left(\sup_{|\boldsymbol{x}-\boldsymbol{y}|\leq\delta} |D^{\boldsymbol{\gamma}} f(\boldsymbol{x}) - D^{\boldsymbol{\gamma}} f(\boldsymbol{y})| \right) \tag{155}$$

Definition A.2 (Best approximation by polynomial). Given a function $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$, its best approximation error by a polynomial of degree k in the compact $K \subseteq \Omega$ is

$$E_k(f) \stackrel{\text{def}}{=} \inf_{p \in P_k^n} \|f - p\|_K \tag{156}$$

In this subsection we will always be using the sup norm for functions, and when it is not clear from the context, we will use the subindex to specify where the supremum is taken. So for example, if f is a function defined in K, $\|f\|_K = \sup_{x \in K} |f(x)|$.

Many results exist on upper bounds to this best approximation error. They are generally called Jackson's theorems (or Jackson's inequalities) in honor to D. Jackson, who first proved that type of results in [22].

The version of Jackson's theorem we will use is the following:

Theorem A.1. *Let $f \in \mathcal{C}_0^m(\mathbb{R}^n)$ with $\operatorname{supp} f \subseteq K \subseteq \mathbb{R}^n$, and K compact, then*

$$E_k(f) \leq C \frac{1}{k^m} \omega_{f,m}\left(\frac{1}{k}\right) \tag{157}$$

where C is a constant depending only on n, m and K.

In our case $K = [-1,1]^n$. Since $f \in \mathrm{W}_m^n$, in particular for any $\boldsymbol{\alpha} \in \mathbb{Z}_{\geq}^n$, $0 \leq |\boldsymbol{\alpha}| \leq m$, $\|D^{\boldsymbol{\alpha}} f\| \leq \sqrt{n}$ and therefore using triangular inequality, for any δ

$$\omega_{f,m}(\delta) \leq 2\sqrt{n} \tag{158}$$

Since C can depend on n, this $2\sqrt{n}$ factor can be added to the constant C, and we directly get the desired result.

A.3.2 Proof of Jackson's Theorem

In [6], the following result is proved:

Theorem A.2. *Let $f \in \mathcal{C}_0^m(\mathbb{R}^n)$ with $\operatorname{supp} f \subseteq K \subseteq \mathbb{R}^n$, K compact, and $\boldsymbol{\alpha} \in \mathbb{Z}_{\geq 0}^n$, then there exists a polynomial p_k of degree at most k in n variables such that*

$$\|D^{\boldsymbol{\alpha}}(f - p_n)\|_K \leq C \frac{1}{k^{m-|\boldsymbol{\alpha}|}} \omega_{f,m-\boldsymbol{\alpha}}\left(\frac{1}{k}\right) \tag{159}$$

We are interested in the case $\boldsymbol{\alpha} = \mathbf{0}$. In the proof we will use some known concepts and results from the theory of Fourier transform.

Definition A.3 (cf. Definition 7.1.2 in [20]). Let the set of functions $\mathscr{S}(\mathbb{R}^n)$ be defined by

$$\mathscr{S}(\mathbb{R}^n) = \{\varphi \in \mathcal{C}^\infty(\mathbb{R}^n) \,:\, \|x^{\boldsymbol{\beta}} D^{\boldsymbol{\alpha}} \varphi\| < \infty \quad \forall \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{Z}_{\geq 0}^n\} \tag{160}$$

Note that $\mathcal{C}_0^\infty(\mathbb{R}^n) \subseteq \mathscr{S}(\mathbb{R}^n)$.

Notation. If $g : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ is a function and $\varepsilon > 0$, $g_{[\varepsilon]}$ is the function defined by

$$g_{[\varepsilon]}(\boldsymbol{x}) = \frac{1}{\varepsilon^n} g\left(\frac{\boldsymbol{x}}{\varepsilon}\right) \tag{161}$$

Also, if $\boldsymbol{z} \in \mathbb{C}^n$, we note its imaginary part as $\operatorname{Im}(\boldsymbol{z}) = (\operatorname{Im}(z_1), \ldots, \operatorname{Im}(z_n))$.

Lemma A.3. *For any m, C, f, δ, the modulus of smoothness of f of order m has the following property:*

$$\omega_{f,m}(C\delta) \leq (C+1)\omega_{f,m}(\delta) \tag{162}$$

PROOF.
Given a multi-integer $\boldsymbol{\gamma}$,

$$\sup_{|\boldsymbol{x}-\boldsymbol{y}|\leq C\delta} |D^{\boldsymbol{\gamma}} f(\boldsymbol{x}) - D^{\boldsymbol{\gamma}} f(\boldsymbol{y})| = \sup_{|\boldsymbol{x}-\boldsymbol{y}|\leq \delta} |D^{\boldsymbol{\gamma}} f(\boldsymbol{x}C) - D^{\boldsymbol{\gamma}} f(\boldsymbol{y}C)| \tag{163}$$

Now the idea is to consider the interval that goes from $\boldsymbol{x}$ to $\boldsymbol{y}$, that are the points $t\boldsymbol{x}+(1-t)\boldsymbol{y}$, for $t \in [0,1]$, and apply triangular inequality for points in that interval. If C is integer, then we can introduce $\sum_{j=1}^{C-1} D^{\boldsymbol{\gamma}} f(C\boldsymbol{x} + (C-j)\boldsymbol{y})$ and apply C times triangular inequality to get

$$|D^{\boldsymbol{\gamma}} f(\boldsymbol{x}C) - D^{\boldsymbol{\gamma}} f(\boldsymbol{y}C)| \leq \sum_{j=0}^{C-1} \left| D^{\boldsymbol{\gamma}} f\Big(j\boldsymbol{x} + (C-j)\boldsymbol{y}\Big) - D^{\boldsymbol{\gamma}} f\Big((j+1)\boldsymbol{x} + (C-(j+1))\boldsymbol{y}\Big)\right| \tag{164}$$

Since for every j, $|j\boldsymbol{x} + (C-j)\boldsymbol{y} - (j+1)\boldsymbol{x} - [C-(j+1)]\boldsymbol{y}| = |\boldsymbol{x}-\boldsymbol{y}|$, and recalling equation (163), for C integer we have:

$$\omega_{f,m}(C\delta) \leq C\omega_{f,m}(\delta) \tag{165}$$

Now if C is not integer, let $\lceil C \rceil$ be the smallest integer which is greater than or equal to C. It is trivially satified:

$$\omega_{f,m}(C\delta) \leq \omega_{f,m}(\lceil C \rceil \delta) \leq \lceil C \rceil \omega_{f,m}(\delta) \leq (C+1)\omega_{f,m}(\delta) \tag{166}$$

■

Lemma A.4. *Let r be a nonnegative integer and $f \in \mathcal{C}_0^r(\mathbb{R}^n)$. For any pair of points $\boldsymbol{x}, \boldsymbol{h} \in \mathbb{R}^n$, the quantity $R(\boldsymbol{x}, \boldsymbol{h})$ defined by*

$$f(\boldsymbol{x}+\boldsymbol{h}) = \sum_{0\leq|\boldsymbol{\alpha}|\leq r} \frac{D^{\boldsymbol{\alpha}} f}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} + R(\boldsymbol{x}, \boldsymbol{h}) \tag{167}$$

where $\boldsymbol{\alpha}! = \prod_{i=1}^n \alpha_i!$ and $\boldsymbol{h}^{\boldsymbol{\alpha}} = \prod_{i=1}^n h_i^{\alpha_i}$, satisfies

$$|R(\boldsymbol{x}, \boldsymbol{h})| \leq \frac{n^r |\boldsymbol{h}|^r \omega_{f,r}(|\boldsymbol{h}|)}{r!} \tag{168}$$

PROOF.
Let us define the function $u : \mathbb{R} \to \mathbb{R}$ by $u(t) \stackrel{\text{def}}{=} f(\boldsymbol{x}+t\boldsymbol{h})$. Applying chain rule for derivatives, the l-th derivative of u is

$$u^{(l}(t) = l! \sum_{|\boldsymbol{\alpha}|=l} \frac{D^{\boldsymbol{\alpha}} f(\boldsymbol{x}+t\boldsymbol{h})}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \tag{169}$$

In addition:

$$\omega_{u,r}(\delta) = \sup_{|t_1-t_2|\leq\delta} \left| u^{(r}(t_1) - u^{(r}(t_2)\right| \tag{170}$$

$$= \sup_{|t_1-t_2|\leq\delta} r! \sum_{|\boldsymbol{\alpha}|=k} \left| \frac{\boldsymbol{h}^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \left[D^{\boldsymbol{\alpha}} f(\boldsymbol{x}+t_1\boldsymbol{h}) - D^{\boldsymbol{\alpha}} f(\boldsymbol{x}+t_2\boldsymbol{h})\right]\right| \qquad \text{By (169)} \tag{171}$$

$$\tag{172}$$

Each term in the sum of (171) is less or equal than the supremum for the first inequality and the supremum is taken only in one of the directions in $\omega_{f,r}$ for the second one we get:

$$\omega_{u,r}(\delta) \leq r!|\boldsymbol{h}|^r \left(\sum_{|\boldsymbol{\alpha}|=k} \frac{1}{\boldsymbol{\alpha}!} \right) \sup_{|t_1\boldsymbol{h}-t_2\boldsymbol{h}|\leq\delta|\boldsymbol{h}|} \sup_{|\boldsymbol{\alpha}|=k} |D^{\boldsymbol{\alpha}} f(\boldsymbol{x}+t_1\boldsymbol{h}) - D^{\boldsymbol{\alpha}} f(\boldsymbol{x}+t_2\boldsymbol{h})| \tag{173}$$

$$\leq r!|\boldsymbol{h}|^r \left(\sum_{|\boldsymbol{\alpha}|=k} \frac{1}{\boldsymbol{\alpha}!} \right) \omega_{f,r}(\delta|h|) \tag{174}$$

$$\tag{175}$$

And finally, using $\sum_{|\boldsymbol{\alpha}|=r} \frac{r!}{\boldsymbol{\alpha}!} = r^n$ we have that

$$\omega_{u,r}(\delta) \leq n^r |\boldsymbol{h}|^r \omega_{f,r}(\delta|\boldsymbol{h}|) \tag{176}$$

Observe that

$$R(\boldsymbol{x},\boldsymbol{h}) = f(\boldsymbol{x}+\boldsymbol{h}) - \sum_{0\leq|\boldsymbol{\alpha}|\leq r} \frac{1}{|\boldsymbol{\alpha}|!}|\boldsymbol{\alpha}|!\frac{D^{\boldsymbol{\alpha}} f(\boldsymbol{x})}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} = u(1) - \sum_{l=1}^{r} \frac{u^{(l}(0)}{l!} \tag{177}$$

Taylor's theorem with the mean-value form of the reminder states that there exists a $\xi \in [0,1]$ such that $u(1) - \sum_{l=1}^{r-1} \frac{u^{(l}(0)}{l!} = \frac{u^{(r}(\xi)}{r!}$. Applying this and (176) with $\delta = 1$ we finish the proof:

$$|R(\boldsymbol{x},\boldsymbol{h})| = \left| \frac{u^{(r}(\xi)}{r!} - \frac{u^{(r}(0)}{r!} \right| \leq \frac{\omega_{u,r}(1)}{r!} \leq \frac{n^r|\boldsymbol{h}|^r\omega_{f,r}(|\boldsymbol{h}|)}{r!} \tag{178}$$

■

Lemma A.5. *Let δ be a fixed positive constant. Then there exists an holomorphic function $G : \mathbb{C}^n \to \mathbb{C}$ and a positive constant A satisfying*

$$|G(\boldsymbol{z})| \leq Ae^{\delta|\mathrm{Im}(\boldsymbol{z})|} \quad \forall \boldsymbol{z} \in \mathbb{C}^n \tag{179}$$

Such that the restriction $g = G|_{\mathbb{R}^n}$ satisfies:

a) $g \in \mathscr{S}(\mathbb{R}^n)$

b) For any integer $r \geq 0$, let

$$I_r = \frac{n^r}{r!} \int_{\mathbb{R}^n} |\boldsymbol{w}|^r(|\boldsymbol{w}|+1)|g(\boldsymbol{w})|d\boldsymbol{w} \tag{180}$$

then for all $f \in \mathcal{C}_0^r(\mathbb{R}^n)$ and $\varepsilon > 0$ it is satisfied that

$$\|f - g_{[\varepsilon]} * f\| \leq I_r\varepsilon^r\omega_{f,r}(\varepsilon) \tag{181}$$

PROOF.
Let $\Phi \in \mathcal{C}_0^\infty(\mathbb{R}^n)$ such that

(*i*) $0 \leq \Phi \leq 1$

(*ii*) There exists an open neighborhood of 0 such that $\Phi = 1$ in it.

(*iii*) $\operatorname{supp}\Phi \subseteq \overline{B(\mathbf{0},\delta)}$ (closed ball of radius δ around the origin)

We define $G(\boldsymbol{z})$ as a Fourier transform of Φ:

$$G(\boldsymbol{z}) \stackrel{\text{def}}{=} \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} \Phi(\boldsymbol{\xi}) e^{-i\boldsymbol{\xi}\cdot\boldsymbol{z}} d\boldsymbol{\xi} \qquad \boldsymbol{z} \in \mathbb{C}^n \tag{182}$$

where $\boldsymbol{x}\cdot\xi$ is the ordinary scalar product. Since Φ has compact support, G is well defined for all $\boldsymbol{z} \in \mathbb{C}^n$. Moreover, the smoothness of Φ lets us exchange derivatives by integrals and gives G is holomorphic in all $\mathbb{C}^n$. Since $\boldsymbol{\xi}$ is real, we can do the following decomposition:

$$e^{-i\boldsymbol{\xi}\cdot\boldsymbol{z}} = e^{-i\boldsymbol{\xi}(\mathrm{Re}(\boldsymbol{z})+i\mathrm{Im}(\boldsymbol{z}))} = e^{\boldsymbol{\xi}\cdot\mathrm{Im}(\boldsymbol{z})} e^{-i\boldsymbol{\xi}\cdot\mathrm{Re}(\boldsymbol{z})} \tag{183}$$

where the second factor has modulus 1. Using this decomposition and the triangular inequality we get:

$$|G(\boldsymbol{z})| \leq \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} \Phi(\boldsymbol{\xi}) e^{\boldsymbol{\xi}\cdot\mathrm{Im}(\boldsymbol{z})} d\boldsymbol{\xi} \tag{184}$$

By property (*iii*) of Φ, (179) is satisfied with $A = \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\Phi(\boldsymbol{\xi})d\boldsymbol{\xi}$. Since $\Phi \in \mathcal{C}_0^\infty(\mathbb{R}^n)$, its Fourier transform is also in $\mathscr{S}(\mathbb{R}^n)$ (this implies $g \in \mathscr{S}(\mathbb{R}^n)$), and applying the Fourier inversion formula [20, Th. 7.1.5]:

$$\Phi(\boldsymbol{\xi}) = \int_{\mathbb{R}^n} g(\boldsymbol{x}) e^{i\boldsymbol{x}\cdot\boldsymbol{\xi}} d\boldsymbol{x} \qquad \boldsymbol{\xi} \in \mathbb{R}^n \tag{185}$$

Note that by setting $\boldsymbol{\xi} = 0$ in (185) and using (*ii*) we get that

$$\int_{\mathbb{R}^n} g(\boldsymbol{x}) d\boldsymbol{x} = 1 \tag{186}$$

If previously we differentiate in (185) with respect to the multi-integer $\boldsymbol{j} = (j_1, \ldots, j_n) \in \mathbb{Z}^n_{\geq 0}$, $\boldsymbol{j} \neq \mathbf{0}$, we get

$$\int_{\mathbb{R}^n} x_1^{j_1} \cdots x_n^{j_n} g(\boldsymbol{x}) d\boldsymbol{x} = 0 \tag{187}$$

We now move to prove property (*b*). For $\boldsymbol{x} \in \mathbb{R}^n$ and $\varepsilon > 0$:

$$(g_{[\varepsilon]} * f - f)(\boldsymbol{x}) = \frac{1}{\varepsilon^n}\int f(\boldsymbol{x}-\boldsymbol{w}) g\left(\frac{\boldsymbol{w}}{\varepsilon}\right) d\boldsymbol{w} - f(\boldsymbol{x}) \tag{188}$$

Applying a change of variables $\boldsymbol{y} = \boldsymbol{w}/\varepsilon$, the first term becomes $\int_{\mathbb{R}^n} f(\boldsymbol{x}-\varepsilon\boldsymbol{y})g(\boldsymbol{y})d\boldsymbol{y}$ and applying (186), so $f(\boldsymbol{x}) = \int_{\mathbb{R}^n} f(\boldsymbol{x})g(\boldsymbol{w})d\boldsymbol{w}$ we get

$$(g_{[\varepsilon]} * f - f)(\boldsymbol{x}) = \int_{\mathbb{R}^n} [f(\boldsymbol{x}-\varepsilon\boldsymbol{w}) - f(\boldsymbol{x})] g(\boldsymbol{w}) d\boldsymbol{w} \tag{189}$$

$$= \int_{\mathbb{R}^n} R(\boldsymbol{x}, \varepsilon\boldsymbol{w}) g(\boldsymbol{w}) d\boldsymbol{w} \qquad \text{Definition of } R(\boldsymbol{x},\boldsymbol{h}) \text{ and (187)} \tag{190}$$

Now using Lemma A.4 and Lemma A.3 we have that

$$|R(\boldsymbol{x}, \varepsilon \boldsymbol{w})| \leq \frac{n^r \varepsilon^r}{r!} \omega_{f,r}(\varepsilon) |\boldsymbol{w}|^r (|\boldsymbol{w}| + 1) \tag{191}$$

from which it directly follows

$$\|g_{[\varepsilon]} * f - f\| \leq \frac{n^r \varepsilon^r}{r!} \omega_{f,r}(\varepsilon) \int_{\mathbb{R}^n} |\boldsymbol{w}^r| (|\boldsymbol{w}| + 1) |g(\boldsymbol{w})| d\boldsymbol{w} \tag{192}$$

which finishes the proof. ■

Definition A.4 (McLaurin polynomials). For any $\boldsymbol{\alpha} \in \mathbb{Z}^n_{\geq 0}$, let $a_{\boldsymbol{\alpha}}$ be its corresponding coefficient on its McLaurin series, so if f is holomorphic in an open neighborhood of 0 fulfills $f(\boldsymbol{z}) = \sum_{\boldsymbol{\alpha}} a_{\boldsymbol{\alpha}} \boldsymbol{z}^{\boldsymbol{\alpha}}$. Then for any nonnegative integer k, the k-th McLaurin polynomial of f is defined as

$$p_{f,k} = \sum_{0 \leq |\boldsymbol{\alpha}| \leq k} a_{\boldsymbol{\alpha}} \boldsymbol{z}^{\boldsymbol{\alpha}} \tag{193}$$

Let $R \geq 0$, we define E_R as the disk of radius R in $\mathbb{C}^n$, namely $E_R = \{\boldsymbol{z} \in \mathbb{C}^n \ : \ |z_i| \leq R \quad \forall i\}$

Lemma A.6. *Let $0 < R < S$. Let f be an holomorphic function in an open neighborhood of E_S satisfying $\|f\|_{E_S} \leq M$. Then*

$$\|f - p_{f,k}\|_{E_R} \leq \frac{M}{1 - R/S} \left(\frac{R}{S}\right)^{k+1} \tag{194}$$

PROOF.
We first prove it for $n = 1$. In that case, let $f(z) = \sum_{\alpha=0}^{\infty} a_\alpha z^\alpha$. Using the definition of a_α and the standard bound for the integral, the hypothesis $\|f\|_{E_S} \leq M$ implies

$$|a_\alpha| = \left| \frac{1}{2\pi i} \oint_{|z|=S} \frac{f(z)}{z^{\alpha+1}} dz \right| \leq \frac{1}{2\pi} \cdot (2\pi S) \frac{M}{S^{\alpha+1}} = \frac{M}{S^\alpha} \tag{195}$$

Now we observe that $(f - p_{f,k})(z) = \sum_{\alpha=k+1}^{\infty} a_\alpha z^\alpha$. Combining both results:

$$\|f - p_{f,k}\|_{E_R} = \sup_{|z| \leq R} \left| \sum_{\alpha=k+1}^{\infty} a_\alpha z^\alpha \right| \tag{196}$$

$$\leq \sup_{|z| \leq R} \sum_{\alpha=k+1}^{\infty} |a_\alpha| |z^\alpha| \qquad \text{Triangular inequality} \tag{197}$$

$$\leq \sum_{\alpha=k+1}^{\infty} \frac{M}{S^\alpha} R^\alpha \qquad \text{(195) and } |z| \leq R \tag{198}$$

$$= \frac{M}{1 - R/S} \left(\frac{R}{S}\right)^{k+1} \qquad \text{Geometric series formula} \tag{199}$$

as required.

For the case of n general, consider a fixed point $\boldsymbol{Z} = (Z_1, \ldots, Z_n) \in \mathbb{C}^n$ such that $|Z_j| \leq 1$ for $j = 1 : n$. Let $\eta : \mathbb{C} \to \mathbb{C}$ be defined by $\eta(\lambda) = f(\lambda \boldsymbol{Z})$. By the chain rule, the k-th derivative of η is

$$\eta^{(k}(\lambda) = \sum_{0 \leq |\boldsymbol{\alpha}| \leq k} \boldsymbol{Z}^{\boldsymbol{\alpha}} D^{\boldsymbol{\alpha}} f(\lambda \boldsymbol{Z}) \tag{200}$$

and because $|Z_j| \leq 1$, the k-th McLaurin coefficient of η is

$$a_k^{\eta} = \frac{1}{2\pi i} \oint_{|z|=S} \frac{\eta^{(k}(\lambda)}{z^{k+1}} \tag{201}$$

and from this it can be derived that $p_{\eta,k}(\lambda) = p_{f,k}(\lambda \boldsymbol{Z})$, so the result for $n = 1$ can be applied. ■

Corollary A.7. *Let $R > 0$, $S > R + 1$ and f be an holomorphic function in an open neighborhood of E_S such that $\|f\|_{E_S} \leq M$. Then*

$$\|f - p_{f,k}\|_{E_R} \leq \frac{M}{1 - R/(S-1)} \left(\frac{R}{S-1} \right)^{k+1} \tag{202}$$

PROOF.
If $S > R + 1$, then $S - 1 > R$, so we can apply Lemma A.6 to $S - 1$ and R instead of S and R, and directly get the result. ■

Lemma A.8. *Let $R > 0$ and $f \in \mathcal{C}^m(\mathbb{R}^n)$ with $\operatorname{supp} f \subseteq [-R, R]^n$. Then for all positive integer k the following inequality holds:*

$$\|f\| \leq R^m (kR + 1) \omega_{f,m} \left(\frac{1}{k} \right) \tag{203}$$

PROOF.
First we prove that

$$\|f\| \leq R^m \sup_{|\boldsymbol{\alpha}|=m} \|D^{\alpha} f\| \tag{204}$$

The case $m = 0$ is obvious. For the case $m = 1$, for each $r \in [-R, R]$, consider the intervals I_r of length R as the intervals $[r - R, r] \times \{0\}^{n-1}$ for $r \in [-R, 0]$ (see Figure 12). Integrating $\partial_{x_1} f$ along I_r we get:

$$\int_{I_r} \partial_{x_1} f = \operatorname{sgn}(r) f(r) \qquad \left| \int_{I_r} \partial_{x_1} f \right| \leq R \sup_{[r-R,r] \times [-R,R]^{n-1}} |\partial_{x_1} f| \tag{205}$$

Making an analogous construction for intervals I_r defined as $[r, r+R] \times \{0\}^{n-1}$ for $r \in [0, R]$, and the analogous construction for the rest of the partial derivatives $\partial_{x_j} f$ for $j = 2 : n$ we get the case $m = 1$. The case for $m > 1$ is proved by applying the case $m = 1$ repeatedly.

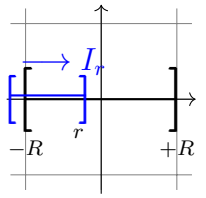

Figure 12: Partial sketch of the proof.

Now by the definition of modulus of smoothness, since $\operatorname{supp} f \subseteq [-R, R]^n$, it is clear that

$$\sup_{|\alpha|=m} \|D^\alpha f\| \leq \omega_{f,m}(R) = \omega_{f,m}\left(k \cdot \frac{R}{k}\right) \leq (kR+1)\omega_{f,m}\left(\frac{1}{k}\right) \tag{206}$$

where the inequality comes from applying Lemma A.3. Equations (204) and (206) together yield to the desired result. ■

PROOF.
(Of Theorem A.2) Let $f \in \mathcal{C}_0^m(\mathbb{R}^n)$. Let R be the diameter of K. Wlog (by a suitable translation in $\mathbb{R}^n$) we can assume $K \subseteq [-R, R]^n$. Let $\delta > 0$ be a fixed real number such that

$$\sqrt{n}(2R+1)\delta < \ln 2 \tag{207}$$

where $\ln 2$ is the natural logarithm of 2.

Let G and $g = G|_{\mathbb{R}^n}$ be the functions of Lemma A.5 associated with δ and for any integer $r \geq 0$, let I_r be as in Lemma A.5. From (207) we see that there exists a constant k_0 such that for all $k \geq k_0$

$$(kR+1)(2Rk)^n e^{\sqrt{n}(2R+1)\delta k} 2^{-k} \leq \frac{I_r}{k^r} \tag{208}$$

Therefore there exists a constant $C \geq I_r$ such that for all $k \geq 1$

$$(kR+1)(2Rk)^n e^{\sqrt{n}(2R+1)\delta k} 2^{-k} \leq \frac{C}{k^r} \tag{209}$$

For the rest of the proof, k will be a fixed positive integer.

Let $H : \mathbb{C}^n \to \mathbb{C}$ be the function such $g_{[\frac{1}{k}]} * f = H|_{\mathbb{R}^n}$,

$$H(\boldsymbol{z}) \stackrel{\text{def}}{=} k^n \int_{\mathbb{R}^n} G\big(k(\boldsymbol{z}-\boldsymbol{w})\big) f(\boldsymbol{w}) d\boldsymbol{w} \qquad \boldsymbol{z} \in \mathbb{C}^n \tag{210}$$

From the above results it follows:

$$\sup_{E_{2R+1}} |H(\boldsymbol{z})| \leq k^n e^{\sqrt{n}(2R+1)\delta k} \int_{\mathbb{R}^n} |f(\boldsymbol{w})| s d\boldsymbol{w} \qquad \text{Lemma A.5} \tag{211}$$

$$\leq A(2Rk)^n R^m (kR+1) e^{\sqrt{n}(2R+1)\delta k} \omega_{f,m}\left(\frac{1}{k}\right) \qquad \text{Lemma A.8} \tag{212}$$

From Lemma A.5 it follows:

$$\sup_{\mathbb{R}^n} |f - g_{[\frac{1}{k}]} * f| \leq \frac{C}{k^m} \omega_{f,m}\left(\frac{1}{k}\right) \tag{213}$$

And from Lemma A.6 with $S = R + 1$ it follows

$$\sup_{[-R,R]^n} |H - p_{H,k}| \leq AR^m(kR+1)(2Rk)^n e^{\sqrt{n}(2R+1)\delta k} 2^{-k} w_{f,m}\left(\frac{1}{k}\right) \tag{214}$$

Now the result comes directly from the concatenation of previous inequalities (214), (213) and (209). ■

A.4 Proof of Theorem 5.1

PROOF.
The left-to-right implication is analogous as in Theorem 5.7. We will focus in the other implication and prove it by contradiction.

We define, for each $\varphi \in \mathcal{C}_0^\infty(\mathbb{R})$, the function $\sigma_\varphi = \sigma * \varphi$.

$$\sigma_\varphi(x) = \int_{-\infty}^{\infty} \sigma(x-y)\varphi(y)dy \tag{215}$$

Since $\sigma, \varphi \in \mathcal{C}(\mathbb{R})$ and φ has compact support, the integral converges for all x. Using Lemma 3.1, $\sigma_\varphi \in \mathcal{C}^\infty(\mathbb{R})$. Taking Riemann sums, $\sigma_\varphi \in \overline{\mathcal{S}_n(\sigma)}$. Using this fact and that

$$\sigma_\varphi(wx+b) = \int_{-\infty}^{\infty} \sigma(wx+b-y)\varphi(y)dy \tag{216}$$

we have that $\overline{\mathcal{S}_n(\sigma_\varphi)} \subseteq \overline{\mathcal{S}_n(\sigma)}$. Because $\sigma_\varphi \in \mathcal{C}^\infty(\mathbb{R})$, we have from the proof of Theorem 5.7, that $x^k \sigma_\varphi^{(k}(b) \in \overline{\mathcal{S}_n(\sigma_\varphi)}$ for all $b \in \mathbb{R}$ and all $k \in \mathbb{N}$.

Suppose $\mathcal{S}_n(\sigma)$ is not dense in $\mathcal{C}(\mathbb{R})$, we will find a contradiction. In that case, there exists k such that $t^k \notin \overline{\mathcal{S}_n(\sigma)}$. Since for each $\varphi \in \mathcal{C}_0^\infty(\mathbb{R})$, $\overline{\mathcal{S}_n(\sigma_\varphi)} \subseteq \overline{\mathcal{S}_n(\sigma)}$, $t^k \notin \overline{\mathcal{S}_n(\sigma_\varphi)}$ for each φ. This implies that $\sigma_\varphi^{(k}(b) = 0$ for all $b \in \mathbb{R}$ and all $\varphi \in \mathcal{C}_0^\infty(\mathbb{R})$. Thus σ_φ is a polynomial of degree at most $k-1$ for each φ. If we set $\varphi_n = \eta_{1/n}$ as in Lemma 3.2, the sequence $\{\sigma_{\varphi_n}\}_{n\in\mathbb{N}}$ tends to σ uniformly (by Lemma 3.2) and by the proof of the left-to-right implication in Theorem 5.7, a sequence of polynomials of degree at most $k-1$, if it converges, the limit is a polynomial of degree at most $k-1$. Thus σ is a polynomial, which is a contradiction. ■