Approximation Capabilities and Complexity Evaluation of Neural ODEs

Jago Silberbauer

Masterarbeit
an der Fakulät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München
Dezember 2022

Betreuer: PD Dr. Dirk - André Deckert

Selbständigkeitserklärung

Hiermit versichere ich, die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel verfasst zu haben. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, den 28.02.2023

_____

Jago Silberbauer

Dedication

This thesis is dedicated to my mother Regina and both my sisters Jana and Nina. Without their continuous and unconditional support in all my endeavors, this work would not have been possible. They are my role models and I am eternally grateful for them.

Acknowledgements

I would like to express my sincerest gratitude and appreciation to my supervisor Dr. Dirk-André Deckert for his guidance, mentorship and valueable advice. This thesis could not have been accomplished without his support, dedication and inspiration.

Contents

1 Introduction

In this thesis, we examine the approximation capabilities of Neural Ordinary Differential Equations depending on the dimension of their input space. Moreover, we argue why a certain Sobolev norm is a reasonable choice for quantifying the richness of the hypothesis space (also called *complexity*) of this model.

In 1989, Hornik ([HSW89]) and Cybenko ([Cyb89]) proved that, for an appropriate non-linear function $\sigma : \mathbb{R} \to \mathbb{R}$ (acting componentwise, if the domain has dimension higher than 1; often called *activation*), and upon choosing $N \in \mathbb{N}$ large enough and finding suitable so-called *weights* $C \in \mathbb{R}^{N\times d}$, $d \in \mathbb{R}^N$, $S \in \mathbb{R}^{m\times N}$, the map

$$K \ni x \mapsto S \cdot \sigma(C \cdot x + d) \in \mathbb{R}^m, \tag{1.1}$$

can approximate any continuous function $h : K \to \mathbb{R}^m$, w.r.t. the $\sup-$norm on K, where $K \subset \mathbb{R}^d$ is compact. The function given in (1.1) is a *neural network with a single hidden layer*, where N is potentially large. *Neural networks with multiple hidden layers* are typically functions $x \mapsto S \cdot y_L[x]$ that are given by the recursion

$$\begin{aligned} K \ni x &\mapsto y_1 := \sigma(C_1 \cdot x + d_1) \\ &\mapsto y_2 := \sigma(C_2 \cdot y_1 + d_2) \\ &\mapsto \dots \\ &\mapsto y_L := \sigma(C_L \cdot y_{L-1} + d_L) \in \mathbb{R}^m, \end{aligned} \tag{1.2}$$

where we choose $L \in \mathbb{N}$ and natural numbers N_l, $l = 0, 1, \dots, L$ with $N_0 = d$. For the weights, we have $C_l \in \mathbb{R}^{N_l \times N_{l-1}}$, $d_l \in \mathbb{R}^{N_l}$ and $S \in \mathbb{R}^{m\times N_L}$ in this case. The number of recursion steps L, or synonymously the number of hidden layers, is called *depth* of the neural network and $N := \max_{l=1,\dots,L} N_l$ its *width*. For such neural networks several approximation results have been proven, where either L or N is unbounded, or even both are; see, e.g., [LPW$^+$17], [KL20], [HSW89] and [Cyb89]. These results are called *universal approximator theorems* and they sparked the idea that neural networks can solve virtually any prediction task. A popular procedure of finding a predictor $x \mapsto p_\theta(x)$ as in (1.2) depending on the collection of all weights θ is the following:

1. Take data $\mathcal{D} := \{(x_j, h(x_j))\}_{j=1,\dots,J}$, $J \in \mathbb{N}$, which represents evaluations of a target function h, i.e., the function that is to be approximated;

2. Choose depth L and $N_1, \dots, N_L$ for your particular network architecture;

3. Define a loss function depending on the weights of the network, where a smaller output means, that the network resembles the data better, e.g.,

$$\mathcal{R}(\theta) := \frac{1}{J} \sum_{j=1}^{J} |h(x_j) - p_\theta(x_j)|^2 ;$$

4. Find weights θ that minimize this loss via, e.g., *Gradient Descent*.

The last step is often called *training* or *learning*. Notice that we have to choose L and N_l, $l = 1, \dots, L$ in the second step. So naturally, questions arise like 'How do we know which depth and width to choose, such that the model performs good on the given data and also generalizes well to unseen inputs?' or 'What consequences arise when the width or depth becomes very large?', to which there are still no satisfactory, general answers.

In 2016, He et al. gave some insights to these questions in [HZRS16]. They describe that the deeper a neural network becomes the more the accuracy of the network during training might degrade. This means that "better training" is not necessarily as simple as increasing the depth of the network in the case of (1.2); in the literature this is sometimes called the "degradation problem".[1] The problem was partially addressed by making use of so-called *Skip Connections*, meaning that an output of a layer is the sum of the previous layer output evaluated by the activation function. Concretely, the recursion in (1.2) with skip connections is recast into

$$\begin{aligned} x \mapsto y_1 &:= x + \sigma(C_1 \cdot x + d_1) \\ &\mapsto y_2 := y_1 + \sigma(C_2 \cdot y_1 + d_2) \\ &\mapsto \dots \\ &\mapsto y_L := y_{L-1} + \sigma(C_L \cdot y_{L-1} + d_L). \end{aligned} \tag{1.3}$$

Such a network is then called a *Residual Neural Network* or *ResNet*[2].

We are not going to be concerned with the degradation problem in this thesis. However, as mathematicians, naturally, we are interested in the infinite layer limit. Namely, heuristically speaking, (1.3) resembles a Euler-Discretization of an Ordinary Differential Equation, i.e., taking the limit $L \to \infty$ of the recursion in (1.3) formally yields an initial value problem of the form

$$\dot{y}(t) = \sigma(C(t) \cdot y(t) + d(t)), \quad y(0) = x, \tag{1.4}$$

where t stems from a finite time horizon, say $t \in [0, 1]$, and the weights are now functions of time namely $C : [0, 1] \to \mathbb{R}^{d \times d}$ and $d : [0, 1] \to \mathbb{R}^d$. We might ask 'Can we create a network architecture with continuous depth that comes from such an ODE?', hoping that it would give more insight to the roles, that depth and width play.

Precisely this was done by Chen et al.([CRBD18]) in 2018. In their paper, a model was formulated in terms of an ODE, where a loss function is optimized w.r.t. the collection of weights Θ, so that the flow of the ODE evaluated at $t = 1$ as a function of the initial datum x, approximates some target function h. Put differently, in the setting of (1.4), functions C and d are found, such that $y(1)$ is close to $h(x)$. Note that here, we have a constant "width" equal to the dimension of the space, from which the initial value stems. This architecture was given the name *Neural Ordinary Differential Equation*.

[1] This passage on the degradation problem is merely here to motivate the model in (1.3). For more details on the matter see [HZRS16].

[2] We need to be careful here, so that the addition in each step is well-defined. For instance we can enforce constant width throughout the layers.

As Neural ODEs seem to have interesting properties that may be advantageous for certain objectives, it is worthwile addressing the fundamental question:

Is there a universal approximator theorem for the model class of

Neural Ordinary Differential Equations?[3]

Or put differently: "Can we prove that Neural ODEs can approximate any continuous function on compact set w.r.t. the sup-norm?". The goal of this thesis is to shine some light on the answers to this question. We are going to see that the width of a Neural ODE plays a crucial role in this. In particular, we give an overview of the complications that arise for this model, in the context of approximation, when the width is too small. Additionally, we show how artificially enlarging width can address these issues.

Contributions

The main contributions of this thesis to the theory of Neural Ordinary Differential Equations can be summarized as follows:

1. Inspired by [DDT19, Section 4], we define a certain class of scalar-valued functions that can not be represented by a Neural ODE composed with a linear transformation. We show in Cor.3.6 that approximation is impossible as well.

2. We show that universal approximation becomes impossible for Neural ODEs in Prop.3.9.

3. We extend a universal approximator theorem for artificially enlarged Neural ODEs proven in [AK20, Section 2] to the case of the output dimension being larger than the input dimension in Thm.3.21.

4. Addressing the problem from the first point above, we prove a general theorem on linear separation via Neural ODEs in Thm.3.23.

5. In Section 5, we argue for properties that a suitable quantifier for the richness of the hypothesis space of Neural ODEs should have. Moreover, we show that a certain Sobolev norm of the parameter function satisfies these desired properties.

[3] In [TTI+20, Section 3], it was proven that composing several Neural ODEs with one another, represented by autonomous ODEs allows for universal approximation of $\mathcal{C}^2$-Diffeomorphisms. In contrast, here, we are concerned with the approximation capabilities of only one Neural ODE that may stem from a non-autonomous ODE.

2 Notations and Definitions

We will now establish notations, definitions and some well-known results that are going to be used in this thesis. Throughout we consider the vector space $\mathbb{R}^d$, $d \in \mathbb{N}$, endowed with the standard scalar product $\langle \cdot, \cdot \rangle$ which induces the euclidean norm, denoted by $|\cdot|$. The topology that is inherited from this norm yields the usual notion of compact sets; for a subset M of a normed space, its closure w.r.t. to a norm $||\cdot||$ is denoted by $\overline{M}^{||\cdot||}$, its boundary by ∂M and its interior is M^{o}. If the closure is taken w.r.t. the euclidean norm, then we simply write $\overline{M}$. Derivatives of a function, say f, at a point x are denoted $f'(x)$ or sometimes df_x; when we derive w.r.t. a time variable we write $\dot{f}$. Integrals in question are always given by the usual Lebesgue-integral. We are also going to make use of the accustomed Landau-symbols o and O whose definitions can be found in [For04].

As mentioned in the introduction, the layers of a neural network depend on so-called *weights*. We symbolize the space of such weights by $\mathcal{W}$. Depending on the type of neural network used, elements of this space might be matrices, vectors, scalars or tuples of such. However, eventually the space of weights is isomorphic to $\mathbb{R}^k$ for some $k \in \mathbb{N}$, and that is what we will work with in this thesis. For a network with L layers (this excludes the input layer but includes the output layer), where $L \in \mathbb{N}$, we need weights for each layer. The space of these weights will be $\mathcal{W}^L := \bigtimes_{l=0}^{L-1} \mathcal{W}$ (one weight for each layer). Note that this definition implies that the network has constant width equal to the dimension of $\mathcal{W}$. On the spaces $\mathcal{W}$ and $\mathcal{W}^L$, we also employ the standard topology.

We now turn to our spaces of *activation functions*. Define $\mathcal{F}(\mathbb{R}^d, \mathcal{W}) \subset \mathcal{C}(\mathbb{R}^d \times \mathcal{W}; \mathbb{R}^d)$ to be the set of continuous functions $f : \mathbb{R}^d \times \mathcal{W} \to \mathbb{R}^d$ for which there exists an increasing function $\gamma_f : [0, \infty[\to [0, \infty[$, such that for all $x, \tilde{x} \in \mathbb{R}^d$ and $\theta \in \mathcal{W}$ we have

$$|f(x, \theta) - f(\tilde{x}, \theta)| \leq \gamma_f(|\theta|) \cdot |x - \tilde{x}|.$$

Note that this is essentially a Lipschitz-estimate in the first argument of f. Moreover, we impose that $\gamma_f(s) = 0$ is a sufficient condition for $s = 0$. Next, we define the subclass of *linearly bounded activation functions* via

$$\mathcal{F}_a(\mathbb{R}^d, \mathcal{W}) := \{f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W}) \mid \exists_{k_a,\, c_a > 0} \forall_{x \in \mathbb{R}^d} \forall_{\theta \in \mathcal{W}} : |f(x, \theta)| \leq k_a \cdot |\theta| + c_a\}.$$

Lastly, we define

$$\mathcal{F}_b(\mathbb{R}^d, \mathcal{W}) := \{f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W}) \mid \exists_{c_b > 0} \forall_{x \in \mathbb{R}^d} \forall_{\theta \in \mathcal{W}} : |f(x, \theta)| \leq c_b\},$$

as the subclass of *bounded activation functions*. As a result, the following inclusions $\mathcal{F}_b(\mathbb{R}^d, \mathcal{W}) \subset \mathcal{F}_a(\mathbb{R}^d, \mathcal{W}) \subset \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ hold.

2.1 Example. Consider the function

$$\sigma : \mathbb{R} \to \mathbb{R}, \quad s \mapsto \sigma(s) := \frac{1}{1 + e^{-s}},$$

known as the *logistic activation function*. Note that the derivative of σ is globally bounded by $1/4$, which, by the mean value theorem, implies that

$$|\sigma(s) - \sigma(\tilde{s})| \leq \frac{1}{4} \cdot |s - \tilde{s}|,$$

for any $s, \tilde{s} \in \mathbb{R}$. We extend this function to $\mathbb{R}^d$, i.e., $\sigma : \mathbb{R}^d \to \mathbb{R}^d$, by letting it act component-wise.

1. Let $\mathcal{W} = \mathbb{R}^d \times \mathbb{R}^{d\times d} \times \mathbb{R}^d \ni \theta = (s, w, b)$. Define the function

$$f_1 : \mathbb{R}^d \times \mathcal{W} \to \mathbb{R}^d, \quad (x, s, w, b) \mapsto f_1(x, s, w, b) := s \odot \sigma(w \cdot x + b),$$

where '$\cdot$' denotes the usual matrix product and '$\odot$' denotes the *Hadamard product* of two vectors, i.e.,

$$s \odot y := (s_1 y_1, \ldots, s_d y_d)$$

for $s, y \in \mathbb{R}^d$. Then, $f_1 \in \mathcal{F}_a(\mathbb{R}^d, \mathcal{W})$.[4] Indeed, it is clear that $f_1 \in \mathcal{C}(\mathbb{R}^d \times \mathcal{W}; \mathbb{R}^d)$ since σ is continuous (the rest is clear from sequential continuity). Furthermore, since all norms are equivalent on $\mathbb{R}^d \times \mathcal{W}$, there exists a $c > 0$ such that for any $x, \tilde{x} \in \mathbb{R}^d$ and any $(s, w, b) \in \mathcal{W}$ the estimate

$$\begin{aligned} |f_1(x, s, w, b) - f_1(\tilde{x}, s, w, b)| &= |s \odot (\sigma(w \cdot x + b) - \sigma(w \cdot \tilde{x} + b))| \\ &\leq c \cdot \max_{k=1,\ldots,d} |s_k(\sigma(\langle w_k, x\rangle + b_k) - \sigma(\langle w_k, \tilde{x}\rangle + b_k)| \end{aligned}$$

holds, where $w_k \in \mathbb{R}^d$ denotes the k-th row of $w \in \mathbb{R}^{d\times d}$. Using the Lipschitz-estimate for σ and the Cauchy-Schwarz-Inequality, this implies

$$\begin{aligned} |f_1(x, s, w, b) - f_1(\tilde{x}, s, w, b)| &\leq \frac{c}{4} \max_{k=1,\ldots,d} |s_k| \cdot |\langle w_k, x - \tilde{x}\rangle| \\ &\leq \frac{c}{4} \left(\max_{k=1,\ldots,d} |s_k| \cdot |w_k| \right) |x - \tilde{x}|. \end{aligned}$$

Lastly, since $|s_k|$ and $|w_k|$ are both less or equal than $|\theta|$, we get

$$|f_1(x, s, w, b) - f_1(\tilde{x}, s, w, b)| \leq \frac{c}{4} |\theta|^2 \cdot |x - \tilde{x}|.$$

Thus, choosing $\gamma_{f_1}(s) := \frac{c}{4} s^2$ yields $f_1 \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$. To see that $f_1 \in \mathcal{F}_a(\mathbb{R}^d, \mathcal{W})$, choose $k_a := c$ and $c_a := 1$. Then, using similar estimates and the fact that σ is globally bounded by 1, it follows that

$$\begin{aligned} |f_1(x, s, w, b| &\leq c \cdot \max_{k=1,\ldots,d} |s_k \sigma(\langle w_k, x\rangle + b_k)| \\ &\leq c \cdot \max_{k=1,\ldots,d} |s_k| \\ &\leq c|\theta| \\ &< k_a|\theta| + c_a. \end{aligned}$$

[4] In this thesis, when we talk about activation functions, we mostly talk about elements of $\mathcal{F}(\mathbb{R}^d, \mathcal{W})$. However, sometimes, like in the present case, the crucial non-linear component is given by some function σ. With an abuse of terminology, we also call this σ an activation function.

2. Let $\mathcal{W} = \mathbb{R}^{d\times d} \times \mathbb{R}^d$, which means that, here, $\theta = (w, b) \in \mathcal{W}$. Define

$$f_2 : \mathbb{R}^d \times \mathcal{W} \to \mathbb{R}^d, \quad (x, w, b) \mapsto f_2(x, w, b) := \sigma(w \cdot x + b).$$

Then, $f_2 \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ follows from analogous estimates as in the prior example. Namely, we can choose $\gamma_{f_2}(s) := \frac{c}{4}s$ and then

$$|f_2(x, \theta) - f_2(\tilde{x}, \theta)| \leq \frac{c}{4}|\theta| \cdot |x - \tilde{x}|$$

holds for any $(x, \theta) \in \mathbb{R}^d \times \mathcal{W}$. Moreover, again since σ is bounded by 1, choosing $c_b := c$, gives

$$|f(x, \theta)| \leq c \cdot \max_{k=1,\dots,d} |\sigma(\langle w_k, x\rangle + b_k)| \leq c = c_b.$$

Hence, $f_2 \in \mathcal{F}_b(\mathbb{R}^d, \mathcal{W})$ holds true.

Note that the choice of σ was somewhat arbitrary, as it just needed to be bounded and have a global Lipschitz-estimate. Many other popular activation functions, such as the *Hyperbolic Tangent* or the *Gaussian* (see Fig.1), satisfy these properties, too.

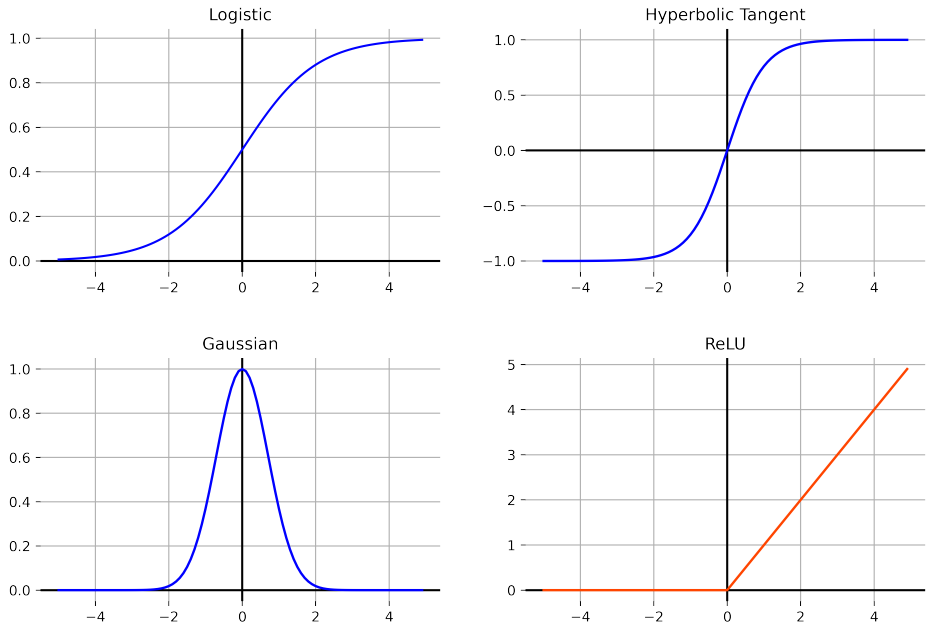

Figure 1: Activation Functions Some examples of popular activation functions $\sigma : \mathbb{R} \to \mathbb{R}$, which are bounded and satisfy a global Lipschitz-estimate, are shown here in blue. However, for an unbounded activation such as the *Rectified Linear Unit* or *ReLU*, shown here in orange, f_1 and f_2 might only belong to $\mathcal{F}(\mathbb{R}^d, \mathcal{W})$ and not necessarily to one of the other subclasses.

In particular, in Section 3, we will talk a lot about convergence of sequences of functions on compact sets w.r.t. to the p-norm and sup-norm (see [LL01] for more information on L^p-spaces). For a compact set $K \subset \mathbb{R}^d$ and a function $h : K \to \mathbb{R}^m$, $m \in \mathbb{N}$, we define the *p-norm of h on K* as

$$||h||_{p,K} := \left(\int_K |h(x)|^p dx \right)^{\frac{1}{p}}$$

and we define the sup-*norm of h on K* as

$$||h||_{\infty,K} := \sup_{x \in K} |h(x)|.$$

We note that due to the compactness of K (implying that it has finite Lebesgue-measure), for any $1 \leq p < p' < \infty$ there exist constants $C, C' > 0$ depending on K such that for all h with finite sup-norm, we have

$$||h||_{p,K} \leq C \cdot ||h||_{p',K} \leq C' \cdot ||h||_{\infty,K}. \tag{2.1}$$

The inequalities in the other direction do not hold in general (see [For12]). The inequality (2.1) tells us that on a compact set K, the convergence w.r.t. $||\cdot||_{p,K}$ gets stronger as p gets larger with $p = \infty$ being the strongest.

Furthermore we will make use of sequences of standard mollifiers, denoted by $(\phi_\eta)_{\eta>0}$. For a definition and a collection of properties see [LL01].

2.1 Neural ODEs and ResNets

We are now ready to introduce Neural ODEs. As illustrated in the introduction, a Neural ODE resembles the flow of an Ordinary Differential Equation, that depends on some *weight function* $\Theta : [0,1] \to \mathcal{W}$. For such a function Θ and $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ we can define the ODE

$$\dot{Z}(t) = f(Z(t), \Theta(t)), \quad t \in]0,1[, \tag{2.2}$$

where $\dot{Z}$ denotes the derivative w.r.t. t. However, for this ODE to be well-defined we need an additional assumption on Θ, namely it needs to be weakly differentiable (see Lem.2.2 for details). Thus, our space of weights for the continuous model will be $\mathcal{W}^\infty := H^1([0,1]; \mathcal{W})$[5], the well-known *Sobolev Space* (see e.g. [LL01]), which is endowed with the norm $||\cdot||_{H^1}$, defined by

$$||\Theta||_{H^1,[0,1]} := ||\Theta||_{2,[0,1]} + ||\dot{\Theta}||_{2,[0,1]}.$$

Here, $\dot{\Theta}$ denotes the weak derivative w.r.t. t.

2.2 Lemma (Well-Posedness of the ODE)**.** *Let $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$, $\Theta \in \mathcal{W}^\infty$ and $x \in \mathbb{R}^d$. Adding the condition $Z(0) = x$, turns* (2.2) *into an initial value problem (IVP). This IVP is well-posed, i.e., a solution $Z \in \mathcal{C}^1([0,1]; \mathbb{R}^d)$ (meaning it is continuously differentiable on $]0,1[$ with a unique continuous extension on $[0,1]$) exists. Moreover, as a solution in $\mathcal{C}^1([0,1]; \mathbb{R}^d)$, it is unique.*

The proof makes use of the well-known theorem due to Picard and Lindelöf and a lemma on the Sobolev norm:

2.3 Theorem (Picard-Lindelöf)**.** *Consider the initial value problem*

$$\begin{aligned} \dot{Z}(t) &= F(t, Z(t)), \quad t \in]0,1[\\ Z(0) &= x, \end{aligned}$$

where $F : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ is globally uniform Lipschitz in the second argument and continuous in the first. Then, there exists a unique solution $Z \in \mathcal{C}^1([0,1]; \mathbb{R}^d)$ to the above ODE that can be extended to the whole of $\mathbb{R}$.

Proof. See [Aul04]. Note that the statement about the extension comes from the fact that the Lipschitz constant is global. □

2.4 Lemma (Sobolev Norm bounds Evaluation)**.** *Let $\Theta \in \mathcal{W}^\infty$. Then, for any $t \in [0,1]$, it holds that*

$$|\Theta(t)| \leq ||\Theta||_{H^1}.$$

Proof. See Appendix A.1. □

[5]There are further reasons for this choice, which are explained in the Sections 4 and 5.

Proof of Lem.2.2. Aiming to use Picard-Lindelöf's theorem, we define for a given $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ and $\Theta \in \mathcal{W}^\infty$ the function

$$F : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d, \quad (t,z) \mapsto F(t,z) := f(z, \Theta(t)).$$

We now show that this function is globally uniform Lipschitz in the second argument and continuous in the first one. The latter is straightforward, as f is continuous and $\Theta \in \mathcal{W}^\infty = H^1([0,1]; \mathcal{W}) \subset \mathcal{C}([0,1]; \mathbb{R}^d)$ thanks to *Sobolev Embeddings* (see [LL01]). We now turn to the desired Lipschitz-estimate. Let $z, \tilde{z} \in \mathbb{R}^d$. Then, using the Lipschitz-property of f, the fact that γ_f is increasing and Lem.2.4, we find for any $t \in [0,1]$

$$\begin{aligned} |F(t,z) - F(t,\tilde{z})| &= |f(z, \Theta(t)) - f(\tilde{z}, \Theta(t))| \\ &\leq \gamma_f(|\Theta(t)|) \cdot |z - \tilde{z}| \\ &\leq \gamma_f(||\Theta||_{H^1}) \cdot |z - \tilde{z}|, \end{aligned}$$

which is precisely the global Lipschitz-property. Applying Thm.2.3 yields the proof. □

We can now properly define what a Neural ODE is:

2.5 Definition (Continuous Flow, Neural ODE)**.** Let $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$. We define the *continuous flow* as the map

$$\begin{aligned} Z_\bullet(\cdot, \star) : [0,1] \times \mathbb{R}^d \times \mathcal{W}^\infty &\to \mathbb{R}^d \\ (t, x, \Theta) &\mapsto Z_t(x, \Theta), \end{aligned} \tag{2.3}$$

where the output is the solution to (2.2) with initial value $x \in \mathbb{R}^d$ and weight function $\Theta \in \mathcal{W}^\infty$ at time $t \in [0,1]$. By Lem.2.2 this map is well-defined. A *Neural ODE* is a function like the one in (2.3), but with fixed weight function Θ and time $t = 1$, i.e.,

$$Z_1(\cdot, \Theta) : \mathbb{R}^d \to \mathbb{R}^d, \quad x \mapsto Z_1(x, \Theta).$$

Furthermore, we denote the set of *Neural ODEs w.r.t. f* by

$$\mathcal{N}_f(\mathbb{R}^d) := \{Z_1(\cdot, \Theta) \mid \Theta \in \mathcal{W}^\infty\}.$$

The dimension of the space from which the initial value for the IVP described in Lem.2.2 stems, is called *width of the Neural ODE*.[6] The length of the time interval on which (2.2) is defined, we call *depth of the Neural ODE*.[7]

Note that for fixed $x \in \mathbb{R}^d$ and $\Theta \in \mathcal{W}^\infty$, the map $Z_\bullet(x, \Theta)$ is a $\mathcal{C}^1$-curve in $\mathbb{R}^d$. We define the length of such a curve as follows:

[6] Note that in this case, the width is equal to the input dimension of the Neural ODE, namely d.

[7] Throughout the thesis we work with depth equal to 1.

2.6 Definition (Length of a Curve)**.** Let $\varphi \in \mathcal{C}^1([0,1];\mathbb{R}^d)$. We define the *length of* φ as

$$l(\varphi) := \int_0^1 |\dot{\varphi}(t)| dt.$$

We will also make use of Grönwall's lemma:

2.7 Lemma (Grönwall's Inequality)**.** *Let $u : [0,1] \to \mathbb{R}$ be continuous and non-negative. Suppose there exist constants $c_u, d_u \geq 0$ such that for any $t \in [0,1]$:*

$$u(t) \leq c_u + d_u \int_0^t u(s) ds.$$

Then, the following estimates hold for all $t \in [0,1]$

$$0 \leq u(t) \leq c_u \cdot e^{d_u t}.$$

Proof. See [Aul04]. Note that in [Aul04] this result only allows for $t \in [0,1[$ in the assumed inequality and in the implied estimate. However, our version then follows immediately from sequential continuity of u. □

Now, we collect some important and well-known properties of the flow of an ODE that we will use throughout the thesis:

2.8 Theorem (Properties of ODE Flows)**.** *Let $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$, $\Theta \in \mathcal{W}^\infty$ and $x, x' \in \mathbb{R}^d$. Then, the following statements hold:*

1. *$Z_0(x, \Theta) = x$*
2. *$Z_\bullet(x, \Theta) \in \mathcal{C}^1([0,1];\mathbb{R}^d)$ and $l(Z_\bullet(x,\Theta)) = \int_0^1 |f(Z_t(x,\Theta), \Theta(t))| dt$*
3. *Trajectories do not intersect: If $x \neq x'$ then for any $t \in [0,1]$ we has $Z_t(x,\Theta) \neq Z_t(x',\Theta)$.*
4. *Lipschitz-continuity in initial data: There exists a constant $C_{f,\Theta} > 0$, such that for any $t \in [0,1]$ on has*
$$|Z_t(x,\Theta) - Z_t(x',\Theta)| \leq C_{f,\Theta} \cdot |x - x'|.$$
5. *Homeomorphism in initial data: For any $t \in [0,1]$ the map $Z_t(\cdot,\Theta)$ is a homeomorphism on $\mathbb{R}^d$.*

Proof. See Appendix A.2. □

So far, we have only seen Neural ODEs as vector fields. However, we will also talk about representing and approximating scalar fields:

2.9 Definition (Scalar Neural ODE)**.** Let $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$. Define the set of *Scalar Neural ODEs* as

$$\mathcal{N}_{f,lin}(\mathbb{R}^d) := \{\mathfrak{L} \circ Z_1(\cdot,\Theta) \mid \mathfrak{L} : \mathbb{R}^d \to \mathbb{R} \text{ linear}, \Theta \in \mathcal{W}^\infty\}.$$

A notion form geometry of great importance to machine learning is *linear separability* of sets. In this thesis, we are going to make statements about a property that is more general:

2.10 Definition ((Homeomorphic) Linear Separability)**.** Let $A, B \subset \mathbb{R}^d$. We say that A and B are *linearly separable* if there exists $y \in \mathbb{R}^d$ and $\omega \in \mathbb{R}$ such that for all $x \in A$ and $x' \in B$ we have

$$\langle y, x \rangle < \omega < \langle y, x' \rangle.$$

We say that A and B are *homeomorphically linear separable* if there exists a homeomorphism $h : \mathbb{R}^d \to \mathbb{R}^d$, such that $h(A)$ and $h(B)$ are linearly separable.

Notice that linear separability implies homeomorphic linear separability, since the identity is a homeomorphism. The converse is not true in general (see Fig.2).

In this thesis, we are mostly concerned with Neural ODEs. However, sometimes taking the discrete perspective of ResNets is helpful. The following lemma, which is a standard result from the theory of numerical approximation of ODEs, makes the relation between the two machine learning architectures clear:

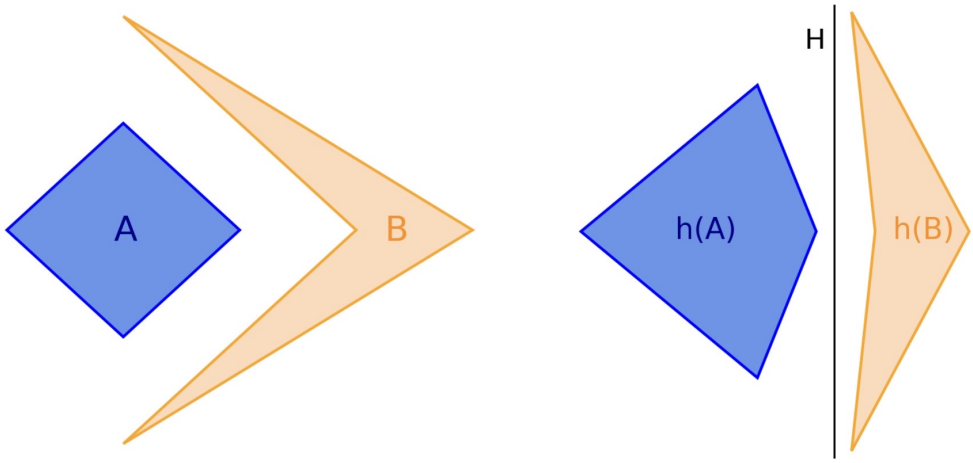

Figure 2: Homeomorphic Linear Separability The sets A and B here are clearly not linearly separable, meaning that it is impossible to find a hyperplane (in this case a straight line) such that A lies entirely on one side of this hyperplane and B on the other side. However, in some cases it might be possible to find a homeomorphism h so that $h(A)$ and $h(B)$ become linearly separable. This can be seen figuratively on the right hand side here. The sets $h(A)$ and $h(B)$ are morphed versions of A and B that can be separated by a hyperplane H.

2.11 Lemma (Convergence of Euler-Discretization)**.** *Let $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W}) \cap \mathcal{C}^2(\mathbb{R}^d \times \mathcal{W}; \mathbb{R}^d)$, $\Theta \in \mathcal{C}^2([0,1]; \mathcal{W}) \subsetneq \mathcal{W}^\infty$ and $K \subset \mathbb{R}^d$ be compact. For any $x \in K$ and $L \in \mathbb{N}$ define the* Euler-Discretization of (f, Θ) *recursively as follows:*

$$\begin{aligned} z_{l+1}[x, \theta^L] &:= z_l[x, \theta^L] + \frac{1}{L} \cdot f(z_l[x, \theta^L], \theta_l^L), \quad l = 0, \dots, L-1 \\ z_0[x, \theta^L] &:= x, \end{aligned} \tag{2.4}$$

where $\theta^L = (\theta_l^L)_{l=0,\dots,L-1} := (\Theta(\frac{l}{L}))_{l=0,\dots,L-1} \in \mathcal{W}^L$. Then, the Euler-Discretization converges to the respective Neural ODE w.r.t. the sup*-norm on K, i.e.,*

$$||z_L[\cdot, \theta^L] - Z_1(\cdot, \Theta)||_{\infty, K} \to 0 \quad as \quad L \to \infty.$$

Proof. Firstly, we consider the special case of $d = 1$. Since $f \in \mathcal{C}^2(\mathbb{R} \times \mathcal{W}; \mathbb{R})$ and $\Theta \in \mathcal{C}^2([0,1]; \mathcal{W})$, we get $Z_\bullet(x, \Theta) \in \mathcal{C}^2([0,1]; \mathbb{R})$ for every $x \in \mathbb{R}$ (see [Aul04]). Using Taylor's theorem and the defining ODE yields

$$\begin{aligned} Z_t(x, \Theta) &= Z_s(x, \Theta) + (t-s) \cdot \dot{Z}_s(x, \Theta) + o(|t-s|^2) \\ &= Z_s(x, \Theta) + (t-s) \cdot f(Z_s(x, \Theta), \Theta(s)) + o(|t-s|^2), \end{aligned} \tag{2.5}$$

for any $s, t, \in [0,1]$. Choosing $t = \frac{l+1}{L}$ and $s = \frac{l}{L}$ implies

$$Z_{\frac{l+1}{L}}(x, \Theta) = Z_{\frac{l}{L}}(x, \Theta) + \frac{1}{L} \cdot f(Z_{\frac{l}{L}}(x, \Theta), \theta_l^L) + o(L^{-2}).$$

For every $l = 0, \dots, L$ and $x \in K$, define $e_l(x) := Z_{\frac{l}{L}}(x, \Theta) - z_l[x, \theta^L]$. By the triangle inequality, equation (2.5) and the Lipschitz-estimate for f, we get

$$|e_{l+1}(x)| \leq |e_l(x)| + \frac{1}{L} \gamma_f(|\theta_l^L|) \cdot |e_l(x)| + o(L^{-2}).$$

We find a constant $C_x > 0$ that depends continuously on x (see [For17] for the mean-value form of the remainder of Taylor-series), such that

$$|e_{l+1}(x)| \leq |e_l(x)| \cdot \left(1 + \frac{1}{L} \gamma_f(|\theta_l^L|)\right) + \frac{C_x}{L^2}.$$

Using the fact that γ_f is increasing, $\theta_l^L = \Theta(\frac{l}{L})$ and Lem.2.4 we obtain

$$|e_{l+1}(x)| \leq |e_l(x)| \cdot \left(1 + \frac{1}{L} \gamma_f(||\Theta||_{H^1})\right) + \frac{C_x}{L^2}.$$

Iterating this inequality yields

$$|e_L(x)| \leq \frac{C_x}{L} \cdot \frac{\left(\frac{1}{L} \cdot \gamma_f(||\Theta||_{H^1}) + 1\right)^L - 1}{\gamma_f(||\Theta||_{H^1})}.$$

Moreover, by well-known properties of Bernoulli's representation for Euler's number, we get

$$|e_L(x)| \leq \frac{C_x}{L} \cdot \frac{e^{\gamma_f(||\Theta||_{H^1})} - 1}{\gamma_f(||\Theta||_{H^1})}.$$

Lastly, as mentioned before, C_x depends continuously on x and since continuous functions on a compact set attain their maximum it follows that

$$||z_L[\cdot, \theta^L] - Z_1(\cdot, \Theta)||_{\infty,K} = \sup_{x \in K} |e_L(x)| \leq \frac{1}{L} \cdot \left(\max_{x \in K} C_x \right) \cdot \frac{e^{\gamma_f(||\Theta||_{H^1})} - 1}{\gamma_f(||\Theta||_{H^1})},$$

which vanishes as $L \to \infty$. This proves the claim. □

Inspired by (2.4), we now define *Residual Neural Networks*:

2.12 Definition/Lemma (Discrete Flow, Residual Neural Network)**.** Let $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ and $L \in \mathbb{N}$. Define the *discrete flow* as the map

$$\begin{aligned} z_\bullet[\cdot, \star] : \{0, \dots, L\} \times \mathbb{R}^d \times \mathcal{W}^L &\to \mathbb{R}^d \\ (l, x, \theta^L) &\mapsto z_l[x, \theta^L], \end{aligned}$$

via the recursion

$$\begin{aligned} z_{l+1}[x, \theta^L] &:= z_l[x, \theta^L] + f(z_l[x, \theta^L], \theta^L_l), \quad l = 0, \dots, L-1 \\ z_0[x, \theta^L] &:= x. \end{aligned} \tag{2.6}$$

This function is well-defined and continuous for all $l = 0, \dots, L$. Moreover, an *L-layer Residual Neural Network / ResNet* is a function given through the above recursion with fixed $\theta^L \in \mathcal{W}^L$ and $l = L$, i.e.,

$$z_L[\cdot, \theta^L] : \mathbb{R}^d \to \mathbb{R}^d, \quad x \mapsto z_L[x, \theta^L].$$

The dimension of the space from which the initial value for the (2.6) stems, is called *width of the ResNet.*[8] The number L is called *depth of the ResNet.*

Proof. See Appendix A.3. □

This ends the section on notations, definitions and well-known results. We continue with a discussion on properties capabilities of Neural ODEs.

[8]Note that in this case, the width is equal to the input dimension of the ResNet, namely d.

3 Approximation Properties

In this section we will examine the approximation properties of Neural ODEs. First, as we will need it later on, we state the original universal approximator theorem for a network with one hidden layer and arbitrarily large width here without proof:

3.1 Theorem ('Vanilla' Neural Networks with 1 Hidden Layer are universal Approximators)**.** *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be continuous with*

$$\lim_{s \to +\infty} \sigma(s) = 1 \quad \textit{and} \quad \lim_{s \to -\infty} \sigma(s) = 0.^{9} \tag{3.1}$$

Moreover, let $K \subset \mathbb{R}^d$ be a compact subset and $h : K \to \mathbb{R}$ a continuous function. Then, for every $\epsilon > 0$ there exists an $N \in \mathbb{N}$ and $s_n, d_n \in \mathbb{R}$, $c_n \in \mathbb{R}^d$, for every $n = 1, \dots, N$, such that the function

$$v : K \to \mathbb{R}, \quad x \mapsto \sum_{n=1}^{N} s_n \sigma(c_n \cdot x + d_n)$$

satisfies $||h - v||_{\infty, K} < \epsilon$.

Proof. See [Cyb89]. □

A straight-forward corollary generalizes this result to functions that have an output space of higher dimension. The proof uses simple Linear Algebra and is retained in the appendix.

3.2 Corollary. *Let $m \leq d$ and $\sigma : \mathbb{R} \to \mathbb{R}$ be continuous with the properties in (.2). Furthermore, let $K \subset \mathbb{R}^d$ be a compact subset, and $h : K \to \mathbb{R}^m$ a continuous function. Then, for every $\epsilon > 0$, there exists an $N \in \mathbb{N}$ and $s_n, d_n \in \mathbb{R}^m$, $C_n \in \mathbb{R}^{m \times d}$, for $n = 1, \dots, N$, such that the function*

$$v : K \to \mathbb{R}^m, \quad x \mapsto \sum_{n=1}^{N} s_n \odot \sigma(C_n \cdot x + d_n)$$

satisfies $||g - v||_{\infty, K} < \epsilon$.

Proof. See Appendix B.1. □

Notice that here, the approximation happens in the width parameter N, i.e., we allow this single hidden layer to be arbitrarily wide. In Subsection 3.1, we will see, that, Neural ODEs cannot approximate w.r.t. the sup-norm. To tackle this problem, we need to "augment the width", which is made precise in Subsection 3.2.

[9] In the literature, this property is often called 'being a *sigmoidal*'.

3.1 Limitations to Approximation

Next, we inspect limitations of (Scalar) Neural ODEs when it comes to approximation. In particular, we will see that universal approximation is impossible. We start with an example of a continuous scalar-valued function, that can not be represented by a Scalar Neural ODE. Ex.3.3 and the presented reasoning therein can be found in [DDT19, Section 4]

3.3 Example (Nested Sets with non-zero distance can not be seperated by a Scalar Neural ODE)**.** Consider a continuous function $g : \mathbb{R}^2 \to \mathbb{R}$ with

$$g\big|_A = -1 \quad \text{and} \quad g\big|_B = +1, \tag{3.2}$$

where $A := \overline{B}_1(0)$ and $B := \overline{B}_3(0) \setminus B_2(0)$. We will see shortly that $g \notin \mathcal{N}_{f,lin}(\mathbb{R}^2)$ for any $f \in \mathcal{F}(\mathbb{R}^2, \mathcal{W})$.

Firstly, we show that A and B are not homeomorphically linear separable. Let $h : \mathbb{R}^2 \to \mathbb{R}^2$ be a homeomorphism. Put $A' := h(A)$, $B' := h(B)$ as well as $C' := h(C)$, where $C := \overline{B}_2(0)$. Since h is a homeomorphism, we get

$$\partial C' = h(\partial C) \quad \text{and} \quad C'^{\mathrm{o}} = h(C^{\mathrm{o}})$$

(see [Arm13]). Moreover, as $\partial C \subset B$, we have $\partial C' \subset B'$. Next, we will see that no subset of C'^{o} is linearly separable from $\partial C'$. Let $D' \subset C'^{\mathrm{o}}$. Again, since h is a homeomorphism, and since C is bounded and connected, so is C'. Any point in the interior of such sets lies on a line, that goes through two points on the boundary $\partial C'$ (see Fig.3). More precisely, for any $x \in D'$ there exist $\lambda \in]0, 1[$ and $y_1, y_2 \in \partial C'$, such that

$$x = \lambda y_1 + (1 - \lambda) y_2. \tag{3.3}$$

Now, suppose that D' and $\partial C'$ are linearly separable, i.e., there exist $\alpha > 0$ and $z \in \mathbb{R}^2$, such that for all $a \in D'$ and $b \in \partial C'$:

$$\langle z, a\rangle < \alpha < \langle z, b\rangle.$$

Together with (3.3), this implies, for any $x \in D'$, that

$$\begin{aligned} \alpha > \langle z, x\rangle &= \langle z, \lambda y_1 + (1 - \lambda) y_2\rangle \\ &= \lambda\langle z, y_1\rangle + (1 - \lambda)\langle z, y_2\rangle \\ &> \lambda\alpha + (1 - \lambda)\alpha \\ &= \alpha, \end{aligned}$$

which is a contradiction. By $A' \subset C'^{\mathrm{o}}$ and $\partial C' \subset B'$ we thus get that A' and B' are not linearly separable.

Suppose further that there exist $f \in \mathcal{F}(\mathbb{R}^2, \mathcal{W})$, $\Theta \in \mathcal{W}^\infty$, and a $v \in \mathbb{R}^2$, such that $g = \langle v, Z_1(\cdot, \Theta)\rangle$. Then, by (3.2), for the images of A and B, we have

$$\langle v, Z_1(A, \Theta)\rangle = \{-1\} \quad \text{and} \quad \langle v, Z_1(B, \Theta)\rangle = \{+1\}.$$

This implies that the sets $Z_1(A, \Theta)$ and $Z_1(B, \Theta)$ are linearly separable. But this is impossible since $Z_1(\cdot, \Theta)$ is a homeomorphism by Thm.2.8 and as we saw above no homeomorphism can make A and B linearly separable. Hence, $g \notin \mathcal{N}_{f,lin}(\mathbb{R}^2)$.

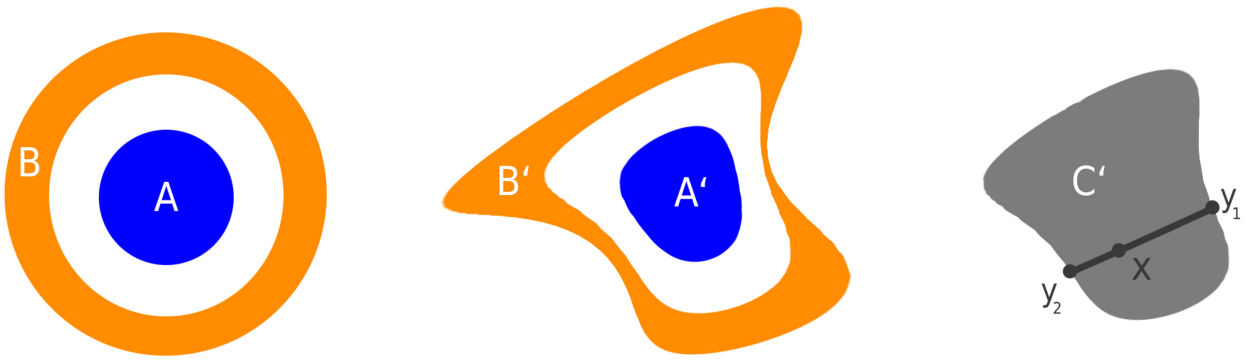

Figure 3: Morphed Nested Sets An exemplary representation of the sets mentioned in 3.3. The sets A and B are not homeomorphically linear separable, since a homeomorphism cannot "tear the outer ring apart".

Inspired by Ex.3.3, we can define a class of scalar functions, that can not be represented by a Scalar Neural ODE, i.e., they do not belong to $\mathcal{N}_{f,lin}(\mathbb{R}^d)$.

3.4 Theorem (Class of unrepresentable Scalar Functions)**.** *Let $g : \mathbb{R}^d \to \mathbb{R}$. Suppose there exist subsets $A, B \subset \mathbb{R}^d$ with*

1. *A and B are not homeomorphically linear seperable and*
2. *there exist $\alpha, \beta \in \mathbb{R}$ with $\alpha < \beta$, such that $g(A) \subset]-\infty, \alpha]$ and $g(B) \subset [\beta, +\infty[$.*

Then, $g \notin \mathcal{N}_{f,lin}(\mathbb{R}^d)$ holds true for any $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$.

Proof. We give a proof by contradiction. Let $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ and suppose $g \in \mathcal{N}_{f,lin}(\mathbb{R}^d)$, i.e., there exist a weight map $\Theta \in \mathcal{W}^\infty$ and a vector $v \in \mathbb{R}^d$, such that $g = \langle v, Z_1(\cdot, \Theta)\rangle$. By *2.* we have

$$\langle v, Z_1(A, \Theta)\rangle \subset]-\infty, \alpha] \quad \text{and} \quad \langle v, Z_1(B, \Theta)\rangle \subset [\beta, +\infty[.$$

This means that for any $z \in Z_1(A, \Theta)$ and any $\tilde{z} \in Z_1(B, \Theta)$ we have that

$$\langle v, z\rangle \leq \alpha < \alpha + \frac{\beta - \alpha}{2} < \beta \leq \langle v, \tilde{z}\rangle,$$

which defines a linear separation of the two sets $Z_1(A, \Theta)$ and $Z_1(B, \Theta)$. However, by *1.* and the fact that the map $Z_1(\cdot, \Theta)$ is a homeomorphism (see Thm.2.8), there cannot be such a separation. We arrive at a contradiction. □

3.5 Example. (Interlocked Double Torus) To give a further example of two sets that are not homeomorphically linear separable, consider the *Interlocked Double Torus* in $\mathbb{R}^3$ which consists of two disjoint sets A and B (see Fig.4). Then, a function $g : \mathbb{R}^3 \to \mathbb{R}$ with say, $g(A) = \{+1\}$ and $g(B) = \{-1\}$, can not be represented by any Scalar Neural ODE.

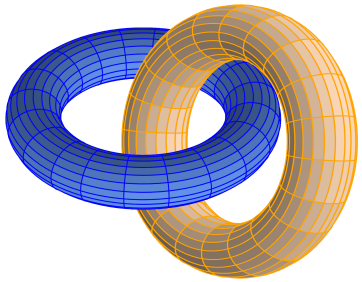

Figure 4: Interlocked Double Torus An example of two sets in $\mathbb{R}^3$ (blue and orange), which are not linearly separable even after applying a homeomorphism, since such cannot "detangle" the two rings.

At this point it is worth noting that, in the literature, the terms *representation* and *approximation* are often used synonymously even though their respective meanings are very much distinct. The first one quantifies which elements belong to $\mathcal{N}_f(\mathbb{R}^d)$ or $\mathcal{N}_{f,lin}(\mathbb{R}^d)$, for some given $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$, while the second does so for the closure of $\mathcal{N}_f(\mathbb{R}^d)$ or $\mathcal{N}_{f,lin}(\mathbb{R}^d)$ w.r.t. some given norm (for us, this is the sup-norm on a compact set). In particular, the latter does not imply the former.

So naturally, a follow-up question is, whether approximation of the scalar functions defined in Thm.3.4 is possible. The next corollary tells us that also this is not feasable.

3.6 Corollary (Class of Scalar Fields that Neural ODEs cannot approximate universally)**.** *Let $g \in \mathcal{C}(\mathbb{R}^d, \mathbb{R})$. Suppose, there exist bounded subsets $A, B \subset \mathbb{R}^d$ with*

1. *A and B are not homeomorphically linearly seperable and*
2. *there exist $\alpha, \beta \in \mathbb{R}$ with $\alpha < \beta$, such that $g(A) \subset]-\infty, \alpha]$ and $g(B) \subset [\beta, +\infty[$.*

Then, $g \notin \overline{\mathcal{N}_{f,lin}(\mathbb{R}^d)}^{||\cdot||_{\infty,K}}$ for any compact $K \supset \overline{A \cup B}$.

Proof. Suppose otherwise, i.e., for any $\epsilon > 0$, there exists $\Theta \in \mathcal{W}^\infty(\mathbb{R}^d)$ and $v \in \mathbb{R}^d$, such that

$$||g - \langle v, Z_1(\cdot, \Theta)\rangle||_{\infty,K} < \epsilon. \tag{3.4}$$

Choose $\epsilon := (\beta - \alpha)/3$ and a corresponding Θ and v so that (3.4) holds. For any $x \in A$

$$|g(x) - \langle v, Z_1(x, \Theta)\rangle| < \epsilon$$

holds true. As $g(A) \subset]-\infty, \alpha]$, by *2.*, we get $\langle v, Z_1(A, \Theta)\rangle \subset]-\infty, \alpha + \epsilon]$. Similarly $\langle v, Z_1(B, \Theta)\rangle \subset [\beta - \epsilon, +\infty[$. However, $\alpha + \epsilon < \beta - \epsilon$ holds and thus,,

$$]-\infty, \alpha + \epsilon] \cap [\beta - \epsilon, +\infty[= \emptyset.$$

Now, we can apply the same argument as in the proof of Thm.3.4 to infer that $Z_1(A, \Theta)$ and $Z_1(B, \Theta)$ must be linearly separable sets in $\mathbb{R}^d$, which contradicts *1.*. This concludes the proof. □

3.7 Remark. Cor.3.6, also implies:

> *"For any compact and connected $K \subset \mathbb{R}^d$, there exists a continuous function $g : K \to \mathbb{R}$, that is not in the closure of $\mathcal{N}_{f,lin}(\mathbb{R}^d)$ w.r.t. $||\cdot||_{\infty,K}$."*.

This is due to the fact that, for any compact and connected K, we can find $r > 0$ small enough, such that the sets

$$A := \{x \in K \mid \text{dist}(x, \partial K) \leq r\} \quad \text{and} \quad B := \{x \in K \mid \text{dist}(x, \partial K) \geq 2r\}$$

are non-empty. Then, we choose a continuous real-valued function g with for example $g(A) = \{+1\}$ and $g(B) = \{-1\}$ and executes the proof as it was demonstrated above.

3.8 Example (Functions with Rotational Symmetry)**.** We give an example of class of functions for which Cor.3.6 applies. Let $g : \mathbb{R}^d \to \mathbb{R}$ such that there exists a continuous and strictly monotone (without loss of generality increasing) function $\psi : [0, \infty[\to \mathbb{R}$ so that for all $x \in \mathbb{R}^d$ we have

$$g(x) = \psi(|x|).$$

We check the assumptions of Cor.3.6. Choose

$$A := \overline{B}_2(0) \setminus B_1(0) \quad \text{and} \quad B := \overline{B}_4(0) \setminus B_3(0).$$

To see that these two sets are not homeomorphically linearly separable we can apply the same argument as in Ex.3.3. For the second condition of Cor.3.6 note that for any $x \in A$ we have $|x| \in [1, 2]$ and thus, by the strict monotonicity of ψ, we have

$$g(x) = \psi(|x|) \in [\psi(1), \psi(2)],$$

where the interval on the right is non-empty. Similarly, $g(x) \in [\psi(3), \psi(4)]$ for $x \in B$. Now, by choosing $\alpha = \psi(2)$ and $\beta = \psi(3)$ we get

$$g(A) \subset [\psi(1), \psi(2)] \subset]-\infty, \alpha] \quad \text{and} \quad g(B) \subset [\psi(3), \psi(4)] \subset [\beta, \infty[.$$

Furthermore, by the strict monotonicity of we get $\alpha < \beta$ and thus, Cor.3.6 can be applied. Hence, $g \notin \overline{\mathcal{N}_{lin}(\mathbb{R}^d)}^{||\cdot||_{\infty,K}}$ for any $K \supset \overline{A \cup B}$. A more specific example of such a function is for instance

$$g : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto |x|^2.$$

So far we have only been talking about scalar-valued functions. We now turn our attention to the topic of vector fields. By lifting up the result from Cor.3.6 we see that universal approximation is impossible for Neural ODEs:

3.9 Proposition (Universal Approximation Fails for Neural ODEs)**.** *Let $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ and $d \geq 1$. Then, for any compact and connected set $K \subset \mathbb{R}^d$, there exists a function $g \in \mathcal{C}(K, \mathbb{R}^d)$, such that*

$$g \notin \overline{\mathcal{N}_f(\mathbb{R}^d)}^{||\cdot||_{\infty,K}}.$$

Proof. Let $K \subset \mathbb{R}^d$ be compact and connected. As explained in Rem.3.7, we find real-valued functions $g_1, \ldots, g_d \in \mathcal{C}(K, \mathbb{R})$, such that for any $i = 1, \ldots, d$ we have

$$g_i \notin \overline{\mathcal{N}_{f,lin}(\mathbb{R}^d)}^{||\cdot||_{\infty,K}}. \tag{3.5}$$

Put $g := (g_1, \ldots, g_d) : K \to \mathbb{R}^d$. This function is continuous, since all of its components are. Our goal is to show that g can not be approximated by a Neural ODE. Suppose this was the case, i.e., for any $\epsilon > 0$ there exists $\Theta_\epsilon \in \mathcal{W}^\infty$, such that

$$||g - Z_1(\cdot, \Theta_\epsilon)||_{\infty,K} < \epsilon.$$

Note that by (3.5) for any $i = 1, \ldots, d$ there exists $\eta_i > 0$, such that for any $\Theta \in \mathcal{W}^\infty$ and $v \in \mathbb{R}^d$ it holds that

$$||g_i - \langle v, Z_1(\cdot, \Theta_\epsilon)\rangle||_{\infty,K} \geq \eta_i.$$

Put $\eta := \min(\eta_1, \ldots, \eta_d)$. Then, by our assumption and the equivalence of norms on finite dimensional vector spaces, for $\epsilon := \frac{\eta}{\sqrt{d+1}}$, we must have the inequalities

$$\begin{aligned}
\frac{\eta}{\sqrt{d+1}} = \epsilon &> ||g - Z_1(\cdot, \Theta_\epsilon)||_{\infty,K} \\
&= \sup_{x \in K} |g(x) - Z_1(x, \Theta_\epsilon)| \\
&\geq \frac{1}{\sqrt{d}} \cdot \sup_{x \in K} \sum_{i=1}^{d} |g_i(x) - \langle e_i, Z_1(x, \Theta_\epsilon)\rangle| \\
&\geq \frac{1}{\sqrt{d}} \cdot \sup_{x \in K} |g_d(x) - \langle e_d, Z_1(x, \Theta_\epsilon)\rangle| \\
&= \frac{1}{\sqrt{d}} \cdot ||g_d - \langle e_d, Z_1(\cdot, \Theta_\epsilon)\rangle||_{\infty,K} \\
&\geq \frac{\eta}{\sqrt{d}}.
\end{aligned}$$

Hence, we arrive at $d + 1 < d$, which is a contradiction. Thus, our assumption cannot hold and we get $g \notin \overline{\mathcal{N}_f(\mathbb{R}^d)}^{||\cdot||_{\infty,K}}$. □

3.10 Remark (The case $d = 1$)**.** The case of $d = 1$ in Prop.3.9 already follows from Thm.3.6. However, in this special case, we can also take a different perspective. We already know that any Neural ODE is a homeomorphism and such have to be strictly monotone for $d = 1$ (otherwise we can yield a contradiction to injectivity together with continuity). Thus, choosing a target function g, that is not injective on some compact set, e.g.,

$$g : \mathbb{R} \to \mathbb{R}, \quad x \mapsto -x^2 + 1,$$

is going to lead to problems, when it comes to approximation w.r.t. the sup-norm on this compactum. More precisely, for $\epsilon > 0$ small enough, approximating g on the compact set $[-1, 1]$ with a Neural ODE yields a contradiction since this Neural ODE would not be injective and hence, not strictly monotone.

It is worth noting shortly that, there is a result by Li et al. ([LLS22]) stating that Neural ODEs are L^p-approximators, i.e., it is possible to approximate any continuous function on a compact set w.r.t. the L^p-norm for $p \geq 1$. However, in this paper $\Theta \in L^\infty([0, 1]; \mathcal{W})$ is allowed. We are not concerned more with this matter in this thesis, but it would certainly be interesting to find out why approximation for $p \geq 1$ is achievable, but for $p = \infty$ it is not (see Prop.3.9).

Summarising, we saw that (Scalar) Neural ODEs face an insurmountable geometric obstacle of not being able to make any two sets linearly separable (as pointed out in Ex.3.3). This is due to the fact that a Neural ODE is always going to be a homeomorphism, and such can only translate between homotopy equivalent sets. As we saw in Cor.3.6 and Prop.3.9, this circumstance leads to classes of functions that can not be approximated by a (Scalar) Neural ODE, which is a severe problem for a machine learning model. Nonetheless, this complication is one of dimensionality, meaning that if we allow our Neural ODEs to act on, say, $\mathbb{R}^{d'}$ for some $d' > d$, we can hope to make any two disjoint sets from $\mathbb{R}^d$ homeomorphically linearly separable in $\mathbb{R}^{d'}$ (with a suitable embedding $\mathbb{R}^d \hookrightarrow \mathbb{R}^{d'}$). In this new scenario, we would have extra dimensions to escape to so that the homeomorphism constraint is still fulfilled. As we are going to see in the next subsection, this in fact works out.

However, before that, we wind up this subsection with another difficulty that Neural ODEs have to face. A Neural ODE is defined via

$$\dot{Z}(t) = f(Z(t), \Theta(t)), \quad t \in]0, 1[,$$

where $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ and $\Theta \in \mathcal{W}^\infty$. Here, the derivative on the left-hand side can be viewed as the velocity in $\mathbb{R}^d$. If we feed an initial value to this ODE, say $x \in \mathbb{R}^d$, the map $[0, 1] \ni t \mapsto Z_t(x, \Theta) \in \mathbb{R}^d$ is the continuous path that is traveled from x to $Z_1(x, \Theta)$ with velocity at time t given by the right-hand side of the equation. However, if we now put a cap on this velocity, say $f \in \mathcal{F}_b(\mathbb{R}^d, \mathcal{W})$, and simultaneously increase the size of the compactum on which we want to approximate, the velocity won't necessarily be enough to reach, say, $g(x)$ within the time window [0,1]. Making this idea more precise is the task of the next theorem.

3.11 Theorem (Universal Approximation fails for bounded Activations)**.** *Let $f \in \mathcal{F}_b(\mathbb{R}^d, \mathcal{W})$ and $d \geq 1$. Then, there exists a compact set $K \subset \mathbb{R}^d$ and a function $g \in \mathcal{C}(K; \mathbb{R}^d)$, such that*

$$g \notin \overline{\mathcal{N}_f(\mathbb{R}^d)}^{||\cdot||_{\infty,K}}.$$

Proof. Since $f \in \mathcal{F}_b(\mathbb{R}^d, \mathcal{W})$, there exists a constant $c_b > 0$ such that for all $x \in \mathbb{R}^d$ and $\theta \in \mathcal{W}$ we have $|f(x, \theta)| \leq c_b$. We put $K := [-N, N]^d$ with $N \in \mathbb{N}$, where the choice of N will be specified later. Moreover, choose $g \in \mathcal{C}(K, \mathbb{R}^d)$ as the function that flips on the hyperplane $\mathbb{R}^{d-1}$ (in the case $d = 1$, flip around the set $\{0\}$). More specifically, for any $x = (x_1, \ldots, x_{d-1}, x_d) \in K$ we have

$$g(x) = (x_1, \ldots, x_{d-1}, -x_d).$$

Note that for $x^* := (0, \ldots, 0, N)$, we easily compute the length of the straight line from x^* to $g(x^*)$ to be

$$|x^* - g(x^*)| = 2N.$$

Our goal is to show that there exists an $\epsilon > 0$, such that for arbitrary $\Theta \in \mathcal{W}^\infty$, we have $||g - Z_1(\cdot, \Theta)||_{\infty,K} \geq \epsilon$. Choose N large and $\epsilon > 0$ small enough, such that $c_b < 2N - \epsilon$. Suppose there exists a $\Theta \in \mathcal{W}^\infty$ such that

$$||g - Z_1(\cdot, \Theta)||_{\infty,K} < \epsilon.$$

This implies $|g(x^*) - Z_1(x^*, \Theta)| < \epsilon$. The shortest path from x^* to $g(x^*)$ in $\mathbb{R}^d$ is the straight line. The map $[0, 1] \ni t \mapsto Z_t(x^*, \Theta) \in \mathbb{R}^d$ is a continuous path starting in x^* and getting ϵ-close to $g(x^*)$. Combining these facts gives rise to the inequalities

$$\begin{aligned} c_b &< 2N - \epsilon \\ &\leq l(Z_\bullet(x^*, \Theta)) \\ &= \int_0^1 |f(Z_t(x^*, \Theta), \Theta(t))| dt \\ &\leq c_b, \end{aligned}$$

which is a contradiction. Thus, no such Θ exists, which is what we had to show. □

An apparent takeaway from Thm.3.11 is that, if we want to pick $f \in \mathcal{F}_b(\mathbb{R}^d, \mathcal{W})$ as an activation function, we should carefully check, whether $c_b > \text{diam}(K)$ holds, where K is the compact set over which the approximation is supposed to happen. Because otherwise, universal approximation becomes impossible on said compactum. Since the size of K is often not known specifically in practice, most of the times we are probably better off picking an element of $\mathcal{F}(\mathbb{R}^d, \mathcal{W}) \setminus \mathcal{F}_b(\mathbb{R}^d, \mathcal{W})$ as an activation function.

3.2 Universal Approximation after Augmenting Width

The previous subsection illustrated the problems that arise for (Scalar) Neural ODEs, when representing/approximating scalar functions and vector fields. Namely, in each case we find functions that cannot be represented/approximated. In this subsection, we see the how augmenting the input dimension for Neural ODEs solves this problem. Def./Lem.3.7, Thm.3.15, Lemmas 3.16 - 3.19 and their respective proofs are all taken from [AK20].

Throughout this subsection, we let $\sigma : \mathbb{R} \to \mathbb{R}$ as in Thm.3.1, i.e., continuous with $\lim_{s\to+\infty} \sigma(s) = 1$ and $\lim_{s\to-\infty} \sigma(s) = 0$. Moreover, we assume σ to be Lipschitz-continuous with constant $L_\sigma > 0$. As before, when we write $\sigma : \mathbb{R}^k \to \mathbb{R}^k$, $k \in \mathbb{N}$, we mean the component-wise extension of σ to $\mathbb{R}^k$. To be able to augment and shrink dimensions we need helping functions:

3.12 Definition (Lift and Projection Function)**.** Let $m \leq d$. Define the *lift function*

$$\mathfrak{l} : \mathbb{R}^d \to \mathbb{R}^{d+m}, \quad x \mapsto \mathfrak{l}(x) := (x, 0),$$

where $0 \in \mathbb{R}^m$. Furthermore, we define the *projection function*

$$\mathfrak{p} : \mathbb{R}^{d+m} \simeq \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^m, \quad (x, y) \mapsto \mathfrak{p}(x, y) := y.$$

Note that these functions depend on the choice of m, however, in the contexts where they are used in, it is usually clear which m was chosen. Moreover, we will make use of the $m \times d$-matrix

$$\mathbb{M} := \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & \dots & 0 \\ \vdots & & \ddots & \ddots & & & \vdots \\ 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{pmatrix},$$

which essentially discards the last $d - m$ components of a vector in $\mathbb{R}^d$.

We are going to show a universal approximation theorem for a certain type of Neural ODE, which we call *Augmented Neural ODE*. It was defined in [AK20], the term however, stems form [DDT19].

3.13 Definition/Lemma (Augmented Neural ODE)**.** Let $m \leq d$ and $\mathcal{W} := \mathbb{R}^{d\times d} \times \mathbb{R}^d \times \mathbb{R}^m$. For any $\Theta = (w, b, s) \in \mathcal{W}^\infty$ and $x \in \mathbb{R}^d$ the following system of ODEs

$$\begin{cases} \dot{\underline{Z}}(t) = w(t)\underline{Z}(t) + b(t), & t \in]0, 1[\\ \dot{\overline{Z}}(t) = s(t) \odot \sigma(\mathbb{M}\underline{Z}(t)), & t \in]0, 1[\\ (\underline{Z}(0), \overline{Z}(0)) = (x, 0) \end{cases} \tag{3.6}$$

is a well-posed initial value problem. Thus, the flow map

$$Z_\bullet(\cdot, \star) : [0, 1] \times \mathbb{R}^{d+m} \times \mathcal{W}^\infty \to \mathbb{R}^{d+m}$$

as well as

$$\mathfrak{p} \circ Z_1(\cdot, \Theta) \circ \mathfrak{l} : \mathbb{R}^d \to \mathbb{R}^m, \quad \text{for } \Theta \in \mathcal{W}^\infty, \tag{3.7}$$

are both well-defined. The latter we call *Augmented Neural ODE*. The dimension of the space from which the initial value for the IVP described in (3.6) stems, is called *width of the Augmented Neural ODE*.[10] The length of the time interval on which (3.6) is defined, we call *depth of the Augmented Neural ODE*.[11][12]

Proof. Define the function

$$f : \mathbb{R}^{d+m} \times \mathcal{W} \to \mathbb{R}^{d+m}, (z, \theta) \mapsto \begin{pmatrix} A\underline{z} + v \\ u \odot \sigma(\mathbb{M}\underline{z}) \end{pmatrix},$$

where $z = (\underline{z}, \overline{z}) \in \mathbb{R}^d \times \mathbb{R}^m \simeq \mathbb{R}^{d+m}$ and $\theta = (A, v, u) \in \mathcal{W} = \mathbb{R}^{d\times d} \times \mathbb{R}^d \times \mathbb{R}^m$. The only thing left to show is $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$; Lem.2.2 then yields the desired well-definedness. Let $z, \tilde{z} \in \mathbb{R}^{d+m}$ and $\theta \in \mathcal{W}$. Since all norms on finite dimensional vector spaces are equivalent, there exists a constant $c > 0$, such that

$$\begin{aligned} |f(z, \theta) - f(\tilde{z}, \theta)| &= \left| \begin{pmatrix} A(\underline{z} - \underline{\tilde{z}}) \\ s \odot (\sigma(\mathbb{M}\underline{z}) - \sigma(\mathbb{M}\underline{\tilde{z}})) \end{pmatrix} \right| \\ &\leq c \cdot (|A| + L_\sigma \cdot |\mathbb{M}| \cdot |s|) \cdot |\underline{z} - \underline{\tilde{z}}| \\ &\leq c(1 + L_\sigma \cdot |\mathbb{M}|) \cdot |\theta| \cdot |z - \tilde{z}|. \end{aligned}$$

Here, we used the Lipschitz-continuity of σ. Now, choosing $\gamma_f(s) := c(1 + L_\sigma \cdot |\mathbb{M}|) \cdot s$ for $s \in [0, \infty[$ gives $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$. □

3.14 Remark (Solution to the System)**.** The system of ODEs given in (3.6) is straightforward to solve. Firstly, note that the right side of the second ODE does not depend on $\overline{Z}$, meaning we can simply integrate it to obtain

$$\overline{Z}(t) = \int_0^t s(r) \odot \sigma(\mathbb{M}\underline{Z}(r))dr, \quad t \in [0, 1].$$

The first equation belongs to a well-known class of ODEs, which can be solved by a method called *Variation of Constants*; see e.g. [For17]. Applying this method yields

$$\underline{Z}(t) = e^{W(t)} \left[\int_0^t e^{-W(s)} b(s)ds + x \right], \quad t \in [0, 1],$$

where $W(t) := \int_0^t w(s)ds$ and we used the well-known matrix exponential. Combining everything yields for input x the following output of an Augmented Neural ODE

$$\overline{Z}(1) = \int_0^1 s(r) \odot \sigma \left(\mathbb{M}e^{W(r)} \left[\int_0^r e^{-W(s)} b(s)ds + x \right] \right) dr.$$

[10] In contrast to the width of the Neural ODE, here, the width is not equal to the input dimension of the Augmented Neural ODE, namely it is $d + m$.

[11] To describe the notation a bit more, notice that for $x \in \mathbb{R}^d$ we have $\mathfrak{p}(Z_1(\mathfrak{l}(x), \Theta)) = \overline{Z}(1)$.

[12] In [AK20], the matrix $\mathbb{M}$ is replaced by a matrix $A \in \mathbb{R}^{m\times d}$ with full rank. For simplicity we stick with $\mathbb{M}$.

With this new setup for Neural ODEs we are now able to prove universal approximation:

3.15 Theorem (Augmented Neural ODEs are Universal Approximators - output dimension small)**.** *Let $h \in \mathcal{C}(K;\mathbb{R}^m)$ with $m \leq d$ and $K \subset \mathbb{R}^d$ compact. Then, for any $\epsilon > 0$ there exist $\Theta = (w, b, s) \in \mathcal{W}^\infty$, such that for any $x \in K$ the Augmented Neural ODE given by* (3.6) *satisfies*

$$|h(x) - \mathfrak{p}(Z_1(\mathfrak{l}(x), \Theta))| < \epsilon.$$

The proof requires four technical lemmas, whose verifications we deposit in Appendix B.2:

3.16 Lemma. *Let $m \leq d$, $s, d \in \mathbb{R}^m$ and $C \in \mathbb{R}^{m\times d}$ with no zero rows. Then, there exist $\tilde{s}_l, \tilde{d}_l \in \mathbb{R}^m$ and $\tilde{C}_l \in \mathbb{R}^{m\times d}$, $l = 1, \ldots, m$, such that*

$$s \odot \sigma(Cx + d) = \sum_{l=1}^{m} \tilde{s}_l \odot \sigma(\tilde{C}_l x + \tilde{d}_l), \tag{3.8}$$

for any $x \in \mathbb{R}^d$, with $\operatorname{rank}(\tilde{C}_l) = m$, for $l = 1, \ldots, m$. If $m = d$ we can choose the matrices $\tilde{C}_l$, such that $\det(\tilde{C}_l) > 0$.

3.17 Lemma. *Let $m \leq d$, $L \in \mathbb{N}$, and $s_l, d_l \in \mathbb{R}^m$ and $C_l \in \mathbb{R}^{m\times d}$ for $l = 1, \ldots, L$ (each C_l having no zero rows). Then, there exists an $\tilde{L} \in \mathbb{N}$ and $\tilde{s}_k, \tilde{d}_k \in \mathbb{R}^m$, $\tilde{C}_k \in \mathbb{R}^{m\times d}$ for $k = 1, \ldots, \tilde{L}$, such that*

$$\sum_{l=1}^{L} s_l \odot \sigma(C_l x + d_l) = \sum_{k=1}^{\tilde{L}} \tilde{s}_k \odot \sigma(\tilde{C}_k x + \tilde{d}_k), \tag{3.9}$$

for any $x \in \mathbb{R}^d$, with $\operatorname{rank}(\tilde{C}_k) = m$, for $k = 0, \ldots, \tilde{L} - 1$. If $m = d$ we can choose the matrices $\tilde{C}_k$, such that $\det(\tilde{C}_k) > 0$.

3.18 Lemma. *Let $m \leq d$. Let $C \in \mathbb{R}^{m\times d}$ with $\operatorname{rank}(C) = m$. If $m = d$ assume additionally $\det(C) > 0$. Then, there exists a $P \in \mathbb{R}^{d\times d}$, such that*

$$C = \mathbb{M}P \quad \textit{and} \quad \det(P) > 0.$$

3.19 Lemma. *Let $p \in [0, \infty[$ and $L \in \mathbb{N}$. For $P_l \in \mathbb{R}^{d\times d}$, $l = 1, \ldots, L$ with $\det(P_l) > 0$ and $0 = t_0 < t_1 < \cdots < t_L = 1$ define the piecewise constant function*

$$P : [0, 1] \to \mathbb{R}^{d\times d}, \quad t \mapsto P(t) := P_L \mathbb{1}_{\{1\}}(t) + \sum_{l=1}^{L} P_l \mathbb{1}_{[t_{l-1}, t_l[}(t).$$

Then, there exists a $c > 0$, such that for any $\eta > 0$, there exists a $P_\eta \in \mathcal{C}^\infty([0, 1]; \mathbb{R}^{d\times d})$ with

$$||P_\eta - P||_{p,[0,1]} < \eta, \quad \det(P_\eta(t)) > 0, \quad |P_\eta(t)| \leq c$$

for all $t \in [0, 1]$.

With these tools we are now prepared to prove universal approximation for Augmented Neural ODEs:

Proof of Thm.3.15. Our strategy is using Cor.3.1 to approximate h with a function v. We then use the above lemmas, to see that we can approximate v with an Augemented Neural ODE. Finally we are going to conclude with the triangle inequality.

For now, suppose $m < d$. Let $h \in \mathcal{C}(K;\mathbb{R}^m)$ and $\epsilon > 0$. By Cor.3.1 we find $N \in \mathbb{N}$ and $s_n, d_n \in \mathbb{R}^m$, $C_n \in \mathbb{R}^{m\times d}$, for $n = 1,\dots,N$, such that the function

$$v : K \to \mathbb{R}^m, \quad x \mapsto \sum_{n=1}^{N} s_n \odot \sigma(C_n \cdot x + d_n)$$

satisfies $||h - v||_{\infty,K} < \frac{\epsilon}{2}$. By Lem.3.17 we can assume without loss of generality, that $\operatorname{rank}(C_n) = m$, for all $n = 1,\dots,N$. Furthermore, by Lem.3.18 there exist matrices $P_n \in \mathbb{R}^{d\times d}$, such that $C_n = \mathbb{M}P_n$ and $\det(P_n) > 0$. Putting $q_n := \mathbb{M}^T d_n$, gives $d_n = \mathbb{M}q_n$, since $\mathbb{M}\mathbb{M}^T = \mathbb{1}$. Moreover, define functions on $[0,1]$ via

$$s(t) := s_n, \quad P(t) := P_n, \quad q(t) := q_n,$$

for $\frac{n-1}{N} \le t < \frac{n}{N}$ and $n = 1,\dots,N$. From this we obtain

$$v(x) = \int_0^1 s(t) \odot \sigma\left(\mathbb{M}(P(t)x + q(t))\right) dt \tag{3.10}$$

for all $x \in K$. Looking at Rem.3.14, this looks similar to the solution of the second equation in (3.6). So our goal now is to approximate the functions s, P and q by smooth functions, whose corresponding integral in the form of (3.10) will be not far off of $v(x)$. Let $(\phi_\eta)_{\eta>0}$ be a sequence of standard mollifiers and put $s_\eta := s * \phi_\eta$, $q_\eta := q * \phi_\eta$, which are smooth functions by well-known results on mollifications (see [LL01]). Additionally, by Lem.3.19, there is a $C > 0$, such that for any $\eta > 0$ there exist smooth functions $P_\eta : [0,1] \to \mathbb{R}^{d\times d}$, such that

$$||P_\eta - P||_{1,[0,1]} < \eta, \quad \det(P_\eta) > 0 \quad \text{and} \quad |P_\eta(t)| \le C$$

for any $t \in [0,1]$. Combining this with facts on mollifications (see [LL01]) we get

$$s_\eta \to s, \quad q_\eta \to q \quad \text{and} \quad P_\eta \to P,$$

as $\eta \to 0$ all w.r.t. the norm $||\cdot||_{1,[0,1]}$. We define our candidate for an Augmented Neural ODE with initial value $(x,0) \in \mathbb{R}^{d\times m}$ to be

$$\begin{aligned} \underline{Z}(t) &:= P_\eta(t)x + q_\eta(t), \\ \overline{Z}(t) &:= \int_0^t s_\eta(r) \odot \sigma(\mathbb{M}\underline{Z}(r))dr. \end{aligned}$$

Two things are left to show now. Firstly, we need $\overline{Z}(1)$ to be $\frac{\epsilon}{2}$-close to $v(x)$ given in (3.10). Secondly, we need to verify, that it can in fact be written in the form an ODE. For the first part observe with the triangle inequality, that

$$|\overline{Z}(1) - v(x)| \leq \int_0^1 |s_\eta(t) \odot \sigma(\mathbb{M}(P_\eta(t)x + q_\eta(t))) - s(t) \odot \sigma\left(\mathbb{M}(P(t)x + q(t))\right)|dt.$$

Adding and subtracting the function $s_\eta \odot \sigma\left(\mathbb{M}(Px + q)\right)$ and using the triangle inequality again yields

$$|\overline{Z}(1) - v(x)| \leq I_1 + I_2,$$

where

$$\begin{aligned} I_1 &:= \int_0^1 |s_\eta(t) - s(t)| \cdot |\sigma\left(\mathbb{M}(P(t)x + q(t))\right)|dt, \\ I_2 &:= \int_0^1 |s_\eta(t)| \cdot |\sigma(\mathbb{M}(P_\eta(t)x + q_\eta(t))) - \sigma\left(\mathbb{M}(P(t)x + q(t))\right)|dt. \end{aligned}$$

Since σ is bounded, by say $M > 0$, we get

$$I_1 \leq M \cdot ||s_\eta - s||_{1,[0,1]}.$$

With properties of the usual standard mollifiers we have

$$\begin{aligned} |s_\eta(t)| &\leq \int_{\mathbb{R}} \phi_\eta(t - r)|s(r)|dr \\ &\leq ||s||_{\infty,[0,1]} \cdot \int_{\mathbb{R}} \phi_\eta(t - r)dr \\ &= ||s||_{\infty,[0,1]}. \end{aligned}$$

Using this together with the Lipschitz continuity of σ we get

$$\begin{aligned} I_2 &\leq ||s||_{\infty,[0,1]} \cdot L_\sigma \cdot |\mathbb{M}| \cdot \int_0^1 |(P_\eta(t) - P(t))x + (q_\eta(t) - q(t))|dt \\ &\leq ||s||_{\infty,[0,1]} \cdot L_\sigma \cdot |\mathbb{M}| \cdot \left(||P_\eta - P||_{1,[0,1]} \cdot \max_{x \in K} |x| + ||q_\eta - q||_{1,[0,1]} \right). \end{aligned}$$

Combining everything we get

$$\begin{aligned} |\overline{Z}(1) - v(x)| \leq\ & M \cdot ||s_\eta - s||_{1,[0,1]} \\ & + ||s||_{\infty,[0,1]} \cdot L_\sigma \cdot |\mathbb{M}| \cdot \left(||P_\eta - P||_{1,[0,1]} \cdot \max_{x \in K} |x| + ||q_\eta - q||_{1,[0,1]} \right), \end{aligned} \tag{3.11}$$

which (independently of x) gets small as η does. Hence, choosing η small enough, such that $|\overline{Z}(1) - v(x)| < \frac{\epsilon}{2}$, gives

$$\begin{aligned} |\overline{Z}(1) - h(x)| &\leq |\overline{Z}(1) - v(x)| + |v(x) - h(x)| \\ &< \epsilon, \end{aligned}$$

for arbitrary $x \in K$. So we are left with showing that our choice of an Augmented Neural ODE actually comes from an ODE. Note that from Lem.3.19 we have that $\det(P_\eta(t)) > 0$ for all $t \in [0,1]$, and thus,, the functions

$$w(t) := \left(\dot{P}_\eta(t)\right) P_\eta(t)^{-1} \quad \text{and} \quad b(t) := \dot{q}_\eta(t) - w(t)q_\eta(t)$$

are well-defined on $[0,1]$. Finally, we find the ODE to be

$$\begin{aligned} \dot{\underline{Z}}(t) &= \dot{P}_\eta(t)x + \dot{q}_\eta(t) \\ &= w(t)P_\eta(t)x + w(t)q_\eta(t) + b(t) \\ &= w(t)\underline{Z}(t) + b(t) \\ \dot{\overline{Z}}(t) &= s_\eta(t) \odot \sigma(\mathbb{M}\underline{Z}(t)), \end{aligned}$$

where $t \in]0,1[$. We have seen that, $\overline{Z}(1)$ approximates v and that it is given via an Augmented Neural ODE.

For the case $m = d$ the procedure is analogous. However, to be able to apply Lem.3.18 and Lem.3.19, we require $\det(C_l) > 0$. But this is already ensured by Lem.3.16 and Lem.3.17, so the proof is finished. □

3.20 Remark (Approximation Speed)**.** From the proof of Thm.3.15 it is clear that the approximation speed of the Augmented Neural ODE can not be better, than the speed of the 1-hidden-layer model. However, we can ask the question, whether it gets worse. Taking a look at (3.11) reveals that the convergence speed of the mollifications towards the functions P, q and s gives an additional restriction on this matter. Namely, we can show that there exists a constant $c > 0$, such that

$$||s_\eta - s||_{1,[0,1]} \leq c \cdot \eta$$

and similarly for P and q (see Appendix B.3 for a precise proof of this statement). Combining, this gives, that the speed of convergence of the Augmented Neural ODE is $O(\epsilon)$ or it's the same as for the 1-hidden-layer model (we have to choose the slower one).

We extend Thm.3.15 to the case of the output dimension being larger than the input dimension:

3.21 Theorem (Augmented Neural ODEs are Universal Approximators - output dimension large)**.** *Let $h \in \mathcal{C}(K;\mathbb{R}^m)$ with $m > d$ and $K \subset \mathbb{R}^d$ compact. Then, for any $\epsilon > 0$ there exist $\Theta = (w, b, s) \in \mathcal{W}^\infty(\mathbb{R}^{m\times m} \times \mathbb{R}^m \times \mathbb{R}^m)$, such that for any $x \in K$, the Augmented Neural ODE given by* (3.7) *satisfies*

$$|h(x) - \mathfrak{p}(Z_1(\mathfrak{l}(x, 0), \Theta))| < \epsilon,$$

where $0 \in \mathbb{R}^{m-d}$.

Proof. Let $\epsilon > 0$. For $h \in \mathcal{C}(K;\mathbb{R}^m)$, consider the function $\tilde{h} : K \times \{0\} \to \mathbb{R}^m$ defined by $\tilde{h}(x, 0) := h(x)$, where $0 \in \mathbb{R}^{m-d}$. Since all of its components are continuous, we also get continuity for $\tilde{h}$, i.e., $\tilde{h} \in \mathcal{C}(K \times \{0\};\mathbb{R}^m)$. Moreover, $K \times \{0\}$ is a compact subset of $\mathbb{R}^m$. Thus, we can apply Thm.3.15 on $\tilde{h}$, meaning there exists $\Theta = (w, b, s) \in \mathcal{W}^\infty(\mathbb{R}^{m\times m} \times \mathbb{R}^m \times \mathbb{R}^m)$, such that for any $(x, 0) \in K \times \{0\}$ we have

$$|\tilde{h}(x, 0) - \mathfrak{p}(Z_1(\mathfrak{l}(x, 0), \Theta))| < \epsilon.$$

By our choice of $\tilde{h}$, this is equivalent to

$$|h(x) - \mathfrak{p}(Z_1(\mathfrak{l}(x, 0), \Theta))| < \epsilon$$

and the proof is finished. □

We conclude this subsection with a result on linear separability. Previously in Cor.3.6, we had the problem of not being able to make certain sets linearly separable with a Neural ODE, due to the fact that such a function is a homeomorphism. Again, augmenting the dimension helps here. To illustrate this, let us look at the following example:

3.22 Example. (Sets that are linearly separable after application of Neural ODE in a higher dimension) We give an example of two sets in $\mathbb{R}$ that are not homeomorphically linearly separable, but become so, after being lifted to $\mathbb{R}^2$. Consider

$$A := [-3, -2] \cup [2, 3] \quad \text{and} \quad B := [-1, 1].$$

In $\mathbb{R}$ a hyperplane corresponds to a single point. However, there is no point in $\mathbb{R}$ that separates the two sets A and B (see Fig.5). Even after applying a homeomorphism $h : \mathbb{R} \to \mathbb{R}$ no such point can be found. This is due to the fact that, as it was mentioned before, homeomorphisms on $\mathbb{R}$ are necessarily strictly monotone and thus,, we get

$$h(A) = [a_1, b_1] \cup [a_3, b_3] \quad \text{and} \quad B := [a_2, b_2],$$

with $a_1 < b_1 < a_2 < b_2 < a_3 < b_3$, which still cannot be separated by a single point. Hence, we know that a Neural ODE is not able to make A and B linearly separable.

To fix this we augment the dimension. Take the lift function $\mathfrak{l} : \mathbb{R} \to \mathbb{R}^2, x \mapsto (x, 0)$ and put $A_0 := l(A)$; B_0 analogously. Moreover, we examine the following system of ODEs:

$$\begin{cases} \dot{z}_1(t) = -z_1(t) \\ \dot{z}_2(t) = z_1(t)^2 + z_2(t), \end{cases}$$

for $t \in]0, 1[$ with some initial value $(z_1(0), z_2(0)) = (x_1, x_2) \in \mathbb{R}^2$. It is not hard to see, that this ODE is well-defined and that the corresponding flow is given by

$$Z_t(x_1, x_2) = \left(x_1 e^{-t}, \left(\frac{x_1^2}{3} + x_2 \right) e^t - \frac{x_1^2}{3} e^{-2t} \right), \quad t \in [0, 1], \quad (x_1, x_2) \in \mathbb{R}^2.$$

Then, from Fig.5 it is clear to see, that A_0 and B_0 are not linearly separable, but $Z_1(A_0)$ and $Z_1(B_0)$ are.

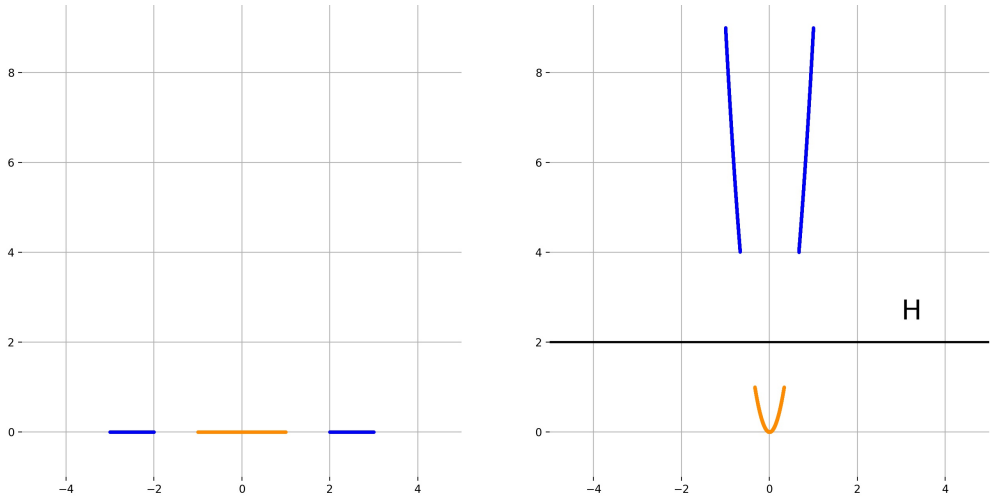

Figure 5: Homeomorphic Linear Separability in higher Dimension *Graph on the left:* The sets A and B are shown in blue and orange respectively. They are not linearly separable. Neither as sets in $\mathbb{R}$, nor as sets in $\mathbb{R}^2$ (i.e., A_0 and B_0). *Graph on the right:* The sets $Z_1(A_0)$ and $Z_1(B_0)$, blue and orange respectively, can be separated by a hyperplane H.

By using Thm.3.15, we can turn the idea of Ex.3.22 into a general result:

3.23 Theorem (Homeomorphic Linear Separability via Neural ODEs after augmenting Dimension)**.** *Let $A, B \subset \mathbb{R}^d$ be bounded and disjoint with*

$$\inf_{a\in A,\, b\in B}|a-b| > 0. \tag{3.12}$$

Put

$$A_0 := A \times \{0\} := \{(a,0) \in \mathbb{R}^{d+1} \mid a \in A\},$$

and B_0 analogously. Then, there exist $f \in \mathcal{F}(\mathbb{R}^{d+1}, \mathcal{W})$ and $\Theta \in \mathcal{W}^\infty$, such that the subsets of $\mathbb{R}^{d+1}$, $Z_1(A_0, \Theta)$ and $Z_1(B_0, \Theta)$, are linearly separable.

Proof. Consider the function $h : \mathbb{R}^d \to \mathbb{R}$ given by $h(x) := (\mathbb{1}_A - \mathbb{1}_B)(x)$ for $x \in \mathbb{R}^d$. Note that $h(A) = \{+1\}$ and $h(B) = \{-1\}$. By well-known density properties of smooth functions with compact support, and the fact that A and B have non-zero distance by (3.12), we can find $\phi \in \mathcal{C}_c^\infty(\mathbb{R}^d; \mathbb{R})$ with

$$||h - \phi||_{\infty,\, \overline{A\cup B}} < \frac{1}{4}.$$

Note that

$$\phi(A) \subset \left]\frac{3}{4}, \frac{5}{4}\right[\quad \text{and} \quad \phi(B) \subset \left]-\frac{5}{4}, -\frac{3}{4}\right[. \tag{3.13}$$

Let

$$\mathfrak{l} : \mathbb{R}^d \to \mathbb{R}^{d+1} \quad \text{and} \quad \mathfrak{p} : \mathbb{R}^{d+1} \to \mathbb{R}^d$$

be the lift and projection function as in Def.3.12. By Thm.3.15 we find $f \in \mathcal{F}(\mathbb{R}^{d+1}, \mathcal{W})$ and $\Theta \in \mathcal{W}^\infty$, such that for any $x \in \overline{A \cup B}$

$$|\phi(x) - \mathfrak{p}\left(Z_1\left(\mathfrak{l}\left(x\right), \Theta\right)\right)| < \frac{1}{4}. \tag{3.14}$$

Note that $A_0 = \mathfrak{l}(A)$ and $B_0 = \mathfrak{l}(B)$. By (3.13) and (3.14) we get for $x \in A$

$$\mathfrak{p}(Z_1(\mathfrak{l}(x), \Theta)) \in \left]\frac{1}{2}, \frac{3}{2}\right[$$

and thus,, $\mathfrak{p}(Z_1(A_0, \Theta)) \subset]\frac{1}{2}, \frac{3}{2}[$. Similarly, we get $\mathfrak{p}(Z_1(B_0, \Theta)) \subset]-\frac{3}{2}, -\frac{1}{2}[$. These two open intervals are disjoint and since $\mathfrak{p}$ is linear we have found a linear separation of the two sets $Z_1(A_0, \Theta)$ and $Z_1(B_0, \Theta)$. This finishes the proof. □

3.24 Remark. The case $\inf_{a\in A,\, b\in B}|a-b| = 0$ results in the boundaries intersecting, i.e., $\partial A \cap \partial B \neq \emptyset$. This yields problems with choosing a suitable $\mathcal{C}_c^\infty$-function in the proof of Thm.3.23, namely, the sup-norm of $h - \phi$ is then going have a lower bound of 1 which is not sufficient to yield the statement. We do not claim that a similar statement in this case is impossible. However, the above proof will not work.

4 Gradient Descent for Neural ODEs

In this thesis, we will not be concerned with the optimization/training aspect of Neural ODEs too much. However, it is still worthwile mentioning, that we can in fact perform a procedure like gradient descent (for an introduction, see e.g. [JS04]) for learning a function $\Theta \in \mathcal{W}^\infty$, which is the goal of this section. We give a setting, where this optimization procedure is well-defined. It will now also become clearer as to why we chose $\mathcal{W}^\infty = H^1([0,1], \mathcal{W})$ as our space of weight functions.

For a given $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$, a *loss function* for a Neural ODE, is a function

$$\mathcal{R}_f : \mathcal{W}^\infty \to [0, \infty[,$$

so that, if $\mathcal{R}_f(\Theta)$, the evaluation at some $\Theta \in \mathcal{W}^\infty$, is small we expect the the corresponding Neural ODE $Z_1(\cdot, \Theta)$ to perform well on some given data $\mathcal{D} = \{(x_j, h(x_j))\}_{j=1,\dots;J}$, $J \in \mathbb{N}$, meaning that we have

$$Z_1(x_j, \Theta) \approx h(x_j).$$

A popular choice for a loss function, given $\mathcal{D}$ and $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$, is

$$\mathcal{R}_f(\Theta) := \frac{1}{J} \sum_{j=1}^{J} |h(x_j) - Z_1(x_j, \Theta)|^2, \tag{4.1}$$

where it is clear, that the closer $\mathcal{R}_f(\Theta)$ gets to 0 the better $Z_1(\cdot, \Theta)$ performs on $\mathcal{D}$. As mentioned in the introduction, we would now like to perform a gradient descent procedure on such a loss function. However, this is a function on an infinite dimensional Banach space. This means that the "gradient" now has to be be a more general form of derivative, namely a *Fréchet derivative*. For a precise definition of this term see Appendix C.1.

The following theorem gives a condition on $\mathcal{R}_f$, under which a method of steepest descent is well-defined:

4.1 Theorem (Gradient Descent on $\mathcal{W}^\infty$)**.** *Let $\mathcal{R} : \mathcal{W}^\infty \to \mathbb{R}$ be Fréchet-differentiable. Then, for any initial $\Theta^{(0)} \in \mathcal{W}^\infty$, the recursive sequence*

$$\Theta^{(k+1)} := \Theta^{(k)} + \eta_k \cdot \Delta^{(k)}, \quad k \in \mathbb{N}_0,$$

with

$$\Delta^{(k)} := \operatorname{argmin}_{||\Delta||_{H^1}=1} d\mathcal{R}_{\Theta^{(k)}}(\Delta) \quad \textit{and} \quad \eta_k := \operatorname{argmin}_{\eta \geq 0} \mathcal{R}\left(\Theta^{(k)} + \eta \cdot \Delta^{(k)}\right),$$

is well-defined.

Note that the theorem does not tell us anything about convergence. For that we might need additional restrictions on $\mathcal{R}$.

Proof. We have to show the existence of $\Delta^{(k)}$ and η_k. This follows from the fact, that $\mathcal{W}^\infty$ is a strictly convex Hilbert space. A detailed proof can be found in [KA16, p. 461ff]. □

For Thm.4.1 to be applicable to a loss function like $\mathcal{R}_f$ in (4.1), we need Fréchet-differentiability of the Neural ODE. The differentiability of $\mathcal{R}_f$ then follows from the chain rule.

4.2 Lemma (Fréchet-Differentiability of Neural ODEs)**.** *Suppose $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ is continuously differentiable (in both variables) with bounded derivative, i.e., there exists an $M > 0$, such that for all $(z, \theta) \in \mathbb{R}^d \times \mathcal{W}$ we have $||df_{(z,\theta)}|| \leq M$.*[13] *Then, for any $t \in [0, 1]$ and $x \in \mathbb{R}^d$, the map*

$$\mathcal{W}^\infty \to \mathbb{R}^d, \quad \Theta \mapsto Z_t(x, \Theta),$$

is Frèchet-differentiable.

Proof. Let $x \in \mathbb{R}^d$ and $t \in]0, 1[$ (the case $t \in \{0, 1\}$ is discussed at the end of the proof). We have to show that for any $\Theta \in \mathcal{W}^\infty$, there exists a bounded linear operator

$$d(Z_t(x, \star))_\Theta : \mathcal{W}^\infty \to \mathbb{R}^d,$$

such that the map

$$\mathcal{W}^\infty \ni \Delta \mapsto |Z_t(x, \Theta + \Delta) - Z_t(x, \Theta) - d(Z_t(x, \star))_\Theta(\Delta)| \in \mathbb{R}$$

is in $o(||\Delta||_{H^1})$. We start by displaying an educated guess for the choice of this bounded operator. In terms of directional derivatives we should have

$$d(Z_t(x, \star))_\Theta(\Delta) = \frac{d}{d\alpha} Z_t(x, \Theta + \alpha \cdot \Delta)\big|_{\alpha=0} =: \phi(\Delta, t).$$

Taking the derivative with respect to t yields another IVP, namely

$$\begin{aligned} \dot{\phi}(\Delta, t) &= df_{(Z_t(x,\Theta),\Theta(t))} \cdot (\phi(\Delta, t), \Delta(t)), \quad t \in]0, 1[\\ \phi(\Delta, 0) &= 0. \end{aligned} \tag{4.2}$$

So our candidate for the Fréchet-derivative at some Θ is going to be

$$d(Z_t(x, \star))_\Theta : \mathcal{W}^\infty \to \mathbb{R}^d, \quad \Delta \mapsto \phi(\Delta, t),$$

where $\phi(\Delta, \bullet)$ is the solution to (4.2). Note that this calculation may be flawed, as it requires an argument for interchanging the derivatives w.r.t. α and t, which was not given. Keep in mind that this is just an educated guess; we are going to see later that this candidate is in fact what we are looking for.

There are now 3 things left to prove. Firstly, well-definedness of this operator. Secondly, we need to verify, that it is in fact linear and bounded. And lastly, we have to show, that

$$u(\Delta, t) := |Z_t(x, \Theta + \Delta) - Z_t(x, \Theta) - \phi(\Delta, t)| \in o(||\Delta||_{H^1}). \tag{4.3}$$

We check these properties in order.

[13] Here $df_{(z,\theta)} : \mathbb{R}^d \times \mathcal{W} \to \mathbb{R}^d$ is a homomorphism and the well-known operator-norm is used. Notice that for a Neural ODE to be well-defined, only continuity was required. For Fréchet differentiability, and in turn the 'training' via gradient descent, we need to impose more on the activation function.

Well-Definedness:

Here we have to show, that the IVP in (4.2) is well-posed for any $\Delta \in \mathcal{W}^\infty$. We are going to use Picard-Lindelöf's theorem. Fix Θ and Δ in $\mathcal{W}^\infty$. Define

$$F : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d, \quad (t, y) \mapsto F(t, y) := df_{(Z_t(x,\Theta),\Theta(t))} \cdot (y, \Delta(t)).$$

Proving, that F is globally uniform Lipschitz in the second argument and continuous in the first, let's us apply Thm.2.3. This yields the desired well-posedness. The latter is clear, since f is continuously differentiable and Δ is continuous by Sobolev-embeddings. For the former, let $y, \tilde{y} \in \mathbb{R}^d$ and observe that

$$\begin{aligned} |F(t, y) - F(t, \tilde{y})| &\leq ||df_{(Z_t(x,\Theta),\Theta(t))}|| \cdot |y - \tilde{y}| \\ &\leq M \cdot |y - \tilde{y}|, \end{aligned}$$

where we used the assumption that the derivative of f is bounded, by some $M > 0$. With this, well-posedness of (4.2) is shown.

Linearity and Boundedness:

We focus on linearity first. Let $\Delta, \tilde{\Delta} \in \mathcal{W}^\infty$ and $s \in \mathbb{R}$. Our goal is to show

$$\phi(s \cdot \Delta + \tilde{\Delta}, t) = s \cdot \phi(\Delta, t) + \phi(\tilde{\Delta}, t). \tag{4.4}$$

Since we now know, that a solution to (4.2) exists and is unique, we can prove the inequality above, by showing, that both sides satisfy the same well-posed ODE. Firstly, note that $\mathcal{W}^\infty$ is a vector space, so $s \cdot \Delta + \tilde{\Delta} \in \mathcal{W}^\infty$. Thus, $\phi(s \cdot \Delta + \tilde{\Delta}, \bullet)$ is well-defined by (4.2). Taking the derivative w.r.t. t of the right-hand side of (4.4) gives

$$\begin{aligned} \frac{d}{dt}(s \cdot \phi(\Delta, t) + \phi(\tilde{\Delta}, t)) &= s \cdot \dot{\phi}(\Delta, t) + \dot{\phi}(\tilde{\Delta}, t) \\ &= s \cdot df_{(Z_t(x,\Theta),\Theta(t))} \cdot (\phi(\Delta, t), \Delta(t)) + df_{(Z_t(x,\Theta),\Theta(t))} \cdot (\phi(\tilde{\Delta}, t), \tilde{\Delta}(t)) \\ &= df_{(Z_t(x,\Theta),\Theta(t))} \cdot (s \cdot \phi(\Delta, t) + \phi(\tilde{\Delta}, t), s \cdot \Delta(t) + \tilde{\Delta}(t)). \end{aligned}$$

Thus, $s \cdot \phi(\Delta, \bullet) + \phi(\tilde{\Delta}, \bullet)$ and $\phi(s \cdot \Delta + \tilde{\Delta}, \bullet)$ satisfy the same well-posed IVP. So by uniqueness, they must be the same. Regarding boundedness, we can use the fundamental theorem of calculus, the triangle-inequality and (4.2) to get

$$\begin{aligned} |\phi(\Delta, t)| &\leq \int_0^t |\dot{\phi}(\Delta, s)| ds \\ &= \int_0^t |df_{(Z_t(x,\Theta),\Theta(t))} \cdot (\phi(\Delta, s), \Delta(s))| ds. \end{aligned}$$

Using the bound on the derivative of f, equivalence of norms on finite-dimensional vector spaces and Lem.2.4 we get that there exists a constant $\tilde{M} > 0$, depending only on f and d, such that

$$\begin{aligned}|\phi(\Delta,t)| &\leq \tilde{M}\cdot\int_0^t|\Delta(s)|ds + \tilde{M}\cdot\int_0^t|\phi(\Delta,s)|ds\\ &\leq \tilde{M}\cdot||\Delta||_{H^1} + \tilde{M}\cdot\int_0^t|\phi(\Delta,s)|ds.\end{aligned}$$

We can now apply Grönwall's Inequality to obtain

$$|\phi(\Delta,t)| \leq \tilde{M}e^{\tilde{M}}\cdot||\Delta||_{H^1},$$

which is the desired boundedness.

Verification of (4.3):

We have to prove that for any $\epsilon > 0$, there exists a $\delta > 0$, such that if $||\Delta||_{H^1} \leq \delta$ we have

$$u(\Delta,t) \leq \epsilon\cdot||\Delta||_{H^1}.$$

So let $\epsilon > 0$. Using the fundamental theorem of calculus, the triangle inequality and the defining ODEs, we obtain

$$\begin{aligned}u(\Delta,t) &\leq \int_0^t|\dot{Z}_s(x,\Theta+\Delta) - \dot{Z}_s(x,\Theta) - \dot{\phi}(\Delta,s)|ds\\ &= \int_0^t|f(Z_s(x,\Theta+\Delta),\Theta(s)+\Delta(s)) - f(Z_s(x,\Theta),\Theta(s))\\ &\qquad - df_{(Z_s(x,\Theta),\Theta(s))}\cdot(\phi(\Delta,s),\Delta(s))|ds.\end{aligned}$$

Notice that

$$\phi(\Delta,s) = Z_s(x,\Theta+\Delta) - Z_s(x,\Theta) - (Z_s(x,\Theta+\Delta) - Z_s(x,\Theta) - \phi(\Delta,s)).$$

By using linearity of $df_{(Z_t(x,\Theta),\Theta(t))}$ and the triangle inequality we get

$$\begin{aligned}u(\Delta,t) \leq\ &\int_0^t\big|f(Z_s(x,\Theta+\Delta),\Theta(s)+\Delta(s)) - f(Z_s(x,\Theta),\Theta(s))\\ &\qquad - df_{(Z_s(x,\Theta),\Theta(s))}\cdot(Z_s(x,\Theta+\Delta) - Z_s(x,\Theta),\Delta(s))\big|ds\\ &+\int_0^t\big|df_{(Z_s(x,\Theta),\Theta(s))}\cdot(Z_s(x,\Theta+\Delta) - Z_s(x,\Theta) - \phi(\Delta,s),0)\big|ds.\end{aligned}$$

The last two integrals we define to be $I_1(t)$ and $I_2(t)$ respectively. We will estimate them separately. For the first one, notice that by properties of the derivative of f, it holds that

$$\begin{aligned}v(\Delta,s) := \big|&f(Z_s(x,\Theta+\Delta),\Theta(s)+\Delta(s)) - f(Z_s(x,\Theta),\Theta(s))\\ &- df_{(Z_s(x,\Theta),\Theta(s))}\cdot(Z_s(x,\Theta+\Delta) - Z_s(x,\Theta),\Delta(s))\big|\end{aligned}$$

is an element of $o(|(Z_s(x,\Theta+\Delta)-Z_s(x,\Theta),\Delta(s))|)$. So for the given $\epsilon>0$, there exists a $\delta'>0$, such that

$$v(\Delta,s)\leq\epsilon\cdot|(Z_s(x,\Theta+\Delta)-Z_s(x,\Theta),\Delta(s))|,$$

whenever $|(Z_s(x,\Theta+\Delta)-Z_s(x,\Theta),\Delta(s))|\leq\delta'$. We further know, that there exists a constant $c>0$, such that

$$|(Z_s(x,\Theta+\Delta)-Z_s(x,\Theta),\Delta(s))|\leq c\cdot||\Delta||_{H^1}.$$

A proof for this inequality can be found in Appendix C.2 (it is yet another application of Grönwall's Inequality). We now choose $\delta:=\delta'/c$. Supposing $||\Delta||_{H^1}\leq\delta$ then yields

$$|(Z_s(x,\Theta+\Delta)-Z_s(x,\Theta),\Delta(s))|\leq\delta'$$

and hence,

$$\begin{aligned}I_1(t)&\leq\epsilon\cdot\int_0^t|(Z_s(x,\Theta+\Delta)-Z_s(x,\Theta),\Delta(s))|ds\\&\leq\epsilon\cdot t\cdot||\Delta||_{H^1}\\&\leq\epsilon\cdot||\Delta||_{H^1}.\end{aligned}$$

For the second integral we use the fact that, the derivative of f is bounded by $M>0$ to obtain

$$I_2(t)\leq M\cdot\int_0^t u(\Delta,s)ds.$$

Combining these two estimates yields

$$u(\Delta,t)\leq\epsilon\cdot||\Delta||_{H^1}+M\cdot\int_0^t u(\Delta,s)ds.$$

Applying Grönwall's Inequality one more time, finally yields

$$u(\Delta,t)\leq\epsilon\cdot||\Delta||_{H^1}\cdot e^M,$$

which is what we had to prove.

We are now left with the case $t\in\{0,1\}$. For $t=0$ we immediately get

$$d(Z_0(x,\star))_\Theta(\Delta)=\frac{d}{d\alpha}x\Big|_{\alpha=0}=0,$$

which is in line with $\phi(\Delta,0)=0$. For $t=1$, we can still apply the fundamental theorem of calculus to $\phi(\Delta,1)$. Thus, the arguments above still work the same, and we get

$$d(Z_1(x,\star))_\Theta(\Delta)=\phi(\Delta,1).$$

This finishes the proof. □

From Thm.4.1 and Lem.4.2 it is now clear, that for the loss function given in (4.1) a Gradient-Descent-Procedure is well-defined. We wrap up this section with a remark on how Neural ODEs are often trained in practice.

4.3 Remark (Adjoint State Method)**.** The way Neural ODEs are used in practice differs from the way they are described mathematically in this thesis. Often the weights are chosen to be independent of t. In paritcular, commonly the right-hand side of (2.2) is a neural network with a set of parameters θ and no time-dependence, making the ODE autonomous.

In this setup we would now like to find "optimal" parameters θ, such that the corresponding flow evaluated at time $t = 1$ is close to some desired target function. Differentiating a flow w.r.t. these parameters is done via the so-called *Adjoint State Method*. A description on how to apply this method can be found in [CLPS03] and [CRBD18].

5 Quantifying Complexity for Neural ODEs

An unwanted trait for models in machine learning or statistics in general is the one of being overfitted. Formally, this means that the model relies too much on the given data and hence, performs poorly on unseen inputs. More precisely, for our case, given a dataset $\mathcal{D} = \{(x_j, y_j)\}_{j=1,\dots,J}$, $J \in \mathbb{N}$, a Neural ODE $Z_1(\cdot, \Theta)$ is *overfitted* to $\mathcal{D}$ if, although the empirical average of

$$|y - Z_1(x, \Theta)|$$

for $(x, y) \in \mathcal{D}$ might be small, the expected value for admissible $(x, y) \notin \mathcal{D}$, is still large. We also say that an overfitted model inherits an "unnecessarily large amount of complexity".

To avoid overfitting during training, via e.g., a gradient descent procedure, we commonly add a so-called *regularization term*, $\mu : \mathcal{W}^\infty \to [0, \infty[$, to the loss function. The regularization term is chosen in a way, such that, during training, models that stem from a needlessly large subset of the set of all possible models get punished. More specifically, for the case of Neural ODEs, μ specifies an increasing sequence of subsets $S_R \uparrow \mathcal{N}_f(\mathbb{R}^d)$, i.e., $S_R \subset S_{R'}$, for $0 \leq R < R'$, and $\bigcup_{R \geq 0} S_R = \mathcal{N}_f(\mathbb{R}^d)$. Now, choosing a model in S_R, with R rather large, should correspond to being able to approximate a larger class of target functions h, while having the risk of ending up with a model that used the extra approximative power to overfit to the given data. We call the evaluation of μ at some weight, the *complexity* of the respective model.

Typical choices for μ are suitable norms on the weight space. That way, weights with large magnitude get punished during training. This results in a simpler model, which is expected to generalize better. In the following we make an argument for the Sobolev norm to be a sensible choice of a regularization term for the Neural ODE model.

5.1 Regularizing Neural ODEs with Sobolev-Norm

Heuristically, we demand three properties from a regularization term / complexity measure μ for our Neural ODEs:

1. *Technical Requirement:*
   The function $\Theta \mapsto \mu(\Theta)$ should be Fréchet-differentiable, so that a Gradient Descent procedure can still be applied (see Thm.4.1).

2. *Restricting Capacity:*
   Allowing only for Θ with $\mu(\Theta) \leq M$, for fixed $M \in]0, \infty[$, limits the approximation capacity of the corresponding Neural ODEs.

3. *Computational Complexity:*
   The depth L of a ResNet that stems from the Neural ODE (in the sense of Lem.2.4) should correspond to $\mu(\Theta)$, in the sense that, $\mu(\Theta)$ determines how large L should be, so that the ResNet, can still achieve a certain accuracy, when approximating a target function h.

In the upcoming three subsections, we shall demonstrate that the following choice of μ lives up to these demnands:

$$\mu : \mathcal{W}^\infty \to \mathbb{R}, \quad \Theta \mapsto \mu(\Theta) := ||\Theta||_{H^1},$$

This μ happens to be a norm on the weight space $\mathcal{W}^\infty$. The increasing sequence of subsets will be

$$S_R := \mathcal{N}_f^R(\mathbb{R}^d) := \{Z_1(\cdot, \Theta) \in \mathcal{N}_f(\mathbb{R}^d) \mid \Theta \in \mu^{-1}([0, R])\},$$

for $R \geq 0$.

5.1.1 Technical Requirement

Our choice of μ is Fréchet differentiable. This stems from the fact, that $\mathcal{W}^\infty = H^1([0,1], \mathcal{W})$ is a Hilbert space. More precisely, it is a well-known result, that norms on Hilbert spaces, which are induced by a scalar product, are Fréchet differentiable everywhere except in 0.

5.1 Lemma (Fréchet-differentiability of $||\cdot||_{H^1}$)**.** *The function*

$$\mu : \mathcal{W}^\infty \to \mathbb{R}, \quad \Theta \mapsto \mu(\Theta) := ||\Theta||_{H^1},$$

is Fréchet differentiable on $\mathcal{W}^\infty \setminus \{0\}$.

Proof. See [Wer06]. □

5.1.2 Restricting Capacity

Next, we would like to see that limiting the complexity of the model should result in restricting the approximation capacity, as explained in the beginning of this section. More specifically, if we fix $R \in [0, \infty[$ and only allow Neural ODEs with weight functions in $\mu^{-1}([0, R])$, then, universal approximation should fail. This is shown for activation functions $f \in \mathcal{F}_a(\mathbb{R}^d, \mathcal{W})$ in the following theorem:

5.2 Theorem (Universal Approximation fails for capped Weight Functions)**.** *Let* $f \in \mathcal{F}_a(\mathbb{R}^d, \mathcal{W})$, $d \geq 1$ *and* $R \in]0, \infty[$. *Then, there exists a compact set* $K \subset \mathbb{R}^d$ *and a function* $g \in \mathcal{C}(K; \mathbb{R}^d)$, *such that*

$$g \notin \overline{\mathcal{N}_f^R(\mathbb{R}^d)}^{||\cdot||_{\infty,K}}.$$

Proof. Since $f \in \mathcal{F}_a(\mathbb{R}^d, \mathcal{W})$, we have for any $\Theta \in \mu^{-1}(]0, R])$, $t \in [0,1]$ and $z \in \mathbb{R}^d$

$$\begin{aligned} |f(z, \Theta(t))| &\leq k_a \cdot |\Theta(t)| + c_a \\ &\leq k_a \cdot ||\Theta||_{H^1} + c_a \\ &\leq k_a \cdot R + c_a. \end{aligned}$$

Here, we used Lem.2.4 in the second estimate. Notice that this essentially means that f is bounded. Recalling the proof of Thm.3.11, we are in the same situation, where both the velocity $f(Z_\bullet(x, \Theta), \Theta(\cdot))$, $x \in \mathbb{R}^d$, and the time-window $[0,1]$ are bounded. From here on out, the proof is completely analogous to the one in Thm.3.11. □

5.3 Remark (Other feasible Bounds on Activation Functions)**.** Notice that we restricted ourselves to activation functions $f \in \mathcal{F}_a(\mathbb{R}^d, \mathcal{W})$ here. However, we could have also allowed any $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ with an estimate of the form

$$|f(z, \theta)| \leq g(|\theta|),$$

where $g : \mathbb{R} \to [0, \infty[$ is continuous. Since any continuous functions attain their maximum on a compact set we then gets

$$|f(z, \theta)| \leq \max_{|\theta| \leq R} g(|\theta|),$$

which also results in boundedness of f. Hence, the proof of Thm.3.11 can be replicated.

5.1.3 Computational Complexity

Lastly, we have the requirement that the complexity μ of a Neural ODE should correspond to the depth L of a corresponding ResNet. This heuristic essentially comes from the triangle inequality:

> *"If a ResNet $z_L[\cdot, \theta]$ can approximate a Neural ODE $Z_1(\cdot, \Theta)$, and $Z_1(\cdot, \Theta)$ can approximate a target function h, then $z_L[\cdot, \theta]$ should be able to approximate h."*

To see this, we merely have to take a look at the inequality (5.1) from Lem.2.11. Namely for any and $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W}) \cap \mathcal{C}^2(\mathbb{R}^d \times \mathcal{W}; \mathbb{R}^d)$ and any compact set $K \subset \mathbb{R}^d$ there exists a constant $C > 0$, such that

$$||z_L[\cdot, \theta^L] - Z_1(\cdot, \Theta)||_{\infty, K} \leq \frac{C}{L} \cdot \frac{e^{\gamma_f(\mu(\Theta))} - 1}{\gamma_f(\mu(\Theta))}. \tag{5.1}$$

From this inequality it is already clear that the right-hand side increases, when $\mu(\Theta)$ does (since γ_f is increasing), and it vanishes, when L becomes sufficiently large. Hence, our third requirement is also fulfilled.

5.4 Remark (L^2 does not quantify Complexity)**.** It is a fair question to ask whether we could have also picked the L^2-norm instead of the H^1-norm as a measure for complexity. The answer is no. This is because Lem.2.4 fails for the L^2-norm, and hence, our requirement on *Computational Complexity* would fail, as we can not obtain an inequality similar to (5.1). More precisely, there exist functions $q \in L^2([0, 1], \mathcal{W})$ such that there is a subset $S \subset [0, 1]$ with non-zero Lebesgue-measure, and

$$||q||_{2,[0,1]} < q(t),$$

for any $t \in S$. For a specific example see Appendix D.1.

5.5 Remark. We conclude this section with a lower bound on the complexity of the Augmented Neural ODE model presented in (3.6). As we saw in the proof of Thm.3.15, for any continuous function h and $\eta > 0$, we can find a $\Theta = (w, b, s) \in \mathcal{W}^\infty$, such that the corresponding Augmented Neural ODE approximates h η-well w.r.t. the sup-norm.

We expect $\mu(\Theta)$ to get larger, as precision η gets smaller, since, heuristically speaking, to achieve a better precision, we have to pay with a higher complexity. We will now prove an estimate that backs this claim. So suppose we want to approximate some continuous target function with an Augmented Neural ODE, specified by $\Theta = (w, b, s) \in \mathcal{W}^\infty$, up to precision $\eta > 0$. Note that we have

$$||\Theta||_{H^1} \geq ||s||_{H^1}.$$

Moreover, consolidating the proof of Thm.3.15, we see that $s = s_\eta$ is a mollification of a step function. Since the Sobolev norm only gets smaller when we restrict to a smaller part of the domain, we can reduce ourselves to the case $s_\eta = \phi_\eta * \mathbb{1}_{[0,1]} : \mathbb{R} \to \mathbb{R}$. We will now prove that there exists a $c > 0$, such that for $\eta > 0$ small enough we have

$$||s_\eta||_{H^1} \geq \frac{c}{\sqrt{\eta}},$$

and hence, after completing the estimate,

$$||\Theta||_{H^1} \geq \frac{c}{\sqrt{\eta}}.$$

Proof of the Complexity Bound: Note that, since $s_\eta \to \mathbb{1}_{[0,1]}$ in L^2 as $\eta \to 0$, by the reverse triangle inequality we have

$$||s_\eta||_2 \to ||\mathbb{1}_{[0,1]}||_2 = 1 \quad \text{as} \quad \eta \to 0.$$

Hence, there exists $\tilde{c} > 0$, s.t. for η small enough $||s_\eta||_2 \geq \tilde{c}$. Moreover a simple computation shows that

$$\begin{aligned} s'_\eta &= \phi'_\eta * \mathbb{1}_{[0,1]} \\ &= \phi_\eta - \phi_\eta(\cdot - 1). \end{aligned}$$

Note that, for η small enough, the summands on the right-hand side here, have disjoint support. Thus we obtain using the transformation formula

$$\begin{aligned} ||s'_\eta||_2^2 &= ||\phi_\eta||_2^2 + ||\phi_\eta(\cdot - 1)||_2^2 \\ &= 2 \cdot ||\phi_\eta||_2^2 \\ &= \frac{2}{\eta} \cdot ||\phi||_2^2. \end{aligned}$$

Combining everything yields

$$\begin{aligned} ||s_\eta||_{H^1} &= ||s_\eta||_2 + ||s'_\eta||_2 \\ &= ||s_\eta||_2 + \sqrt{\frac{2}{\eta}} \cdot ||\phi||_2 \\ &\geq \min(\tilde{c}, \sqrt{2} \cdot ||\phi||_2) \cdot \left(1 + \frac{1}{\sqrt{\eta}}\right) \\ &\geq \frac{c}{\sqrt{\eta}} \end{aligned}$$

for η small enough, where we chose $c := \min(\tilde{c}, \sqrt{2} \cdot ||\phi||_2)$. □

Conclusion

We saw that no universal approximation theorem for (Scalar) Neural ODEs can exist in Cor.3.6 and Prop.3.6. This is due to the fact that a Neural ODE is always a homeomorphism, and such preserve the topology of the input space. However, if a target function does not have this property, representation as well as approximation is impossible. Nevertheless, it is manageable to tackle this problem by augmenting the dimension of the input space. After this modification of the Neural ODE model, there are no obsticles, and universal approximation can be achieved (even for larger output dimensions as we saw in Thm.3.21). Moreover, in Thm.3.23, we proved that any two bounded disjoint subset from $\mathbb{R}^d$ with non-zero distance can be turned into linearly separable sets by applying an Augmented Neural ODE.

Furthermore, we showed that choosing the Sobolev space H^1 as the space of weight functions allows us to define a Gradient Descent Procedure. Here, in Lem.4.2, we gave a condition on activation functions $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ for which Neural ODEs become Fréchet differentiable w.r.t. the weight function.

Lastly, we formulated three sensible criteria for a regularization term for Neural ODEs. We proved that the Sobolev norm satisfies these specifications and is, hence, a reasonable quantifier for complexity.

Outlook

The result [AK20, Theorem 2.3] shows that a width for Neural ODEs equal to input dimension plus output dimension is sufficient to achieve universal approximation. However, it is not clear, whether a smaller width is also sufficient for certain classes of target functions and it would be interesting to explore this further.

Moreover, as we saw in Section 3.1, homeomorphic linear separability is a crucial attribute of sets when it comes to modeling strengths of Neural ODEs. Characterizing this property in a more practical sense may help to specify further for which tasks Neural ODEs are a sensible choice for a modle.

Finally, the idea for the Neural ODE model was inspired by the Euler discretization of ODEs. A next step could be to consider the "Euler-Maruyama discretization" and make sense of *Neural Stochastic Differential Equations*. E.g., we can again ask the question, whether universal approximation is possible.

References

[AK20] Yuto Aizawa and Masato Kimura. Universal Approximation Properties for ODENet and ResNet. *arXiv preprint arXiv:2101.10229*, 2020.

[Arm13] Mark Anthony Armstrong. *Basic Topology*. Springer Science & Business Media, 2013.

[Aul04] Bernd Aulbach. *Gewöhnliche Gifferentialgleichungen*. Springer Spektrum, 2004.

[Bak03] Andrew Baker. *Matrix Groups: An Introduction to Lie Group Theory*. Springer Science & Business Media, 2003.

[CLPS03] Yang Cao, Shengtai Li, Linda Petzold, and Radu Serban. Adjoint Sensitivity Analysis for Differential-Algebraic Equations: The Adjoint DAE System and Its Numerical Solution. *SIAM journal on scientific computing*, 24(3):1076–1089, 2003.

[CRBD18] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural Ordinary Differential Equations. *Advances in neural information processing systems*, 31, 2018.

[Cyb89] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[DDT19] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented Neural ODEs. *Advances in Neural Information Processing Systems*, 32, 2019.

[For04] Otto Forster. *Analysis 1: Differential- und Integralrechnung einer Veränderlichen*. Springer, 2004.

[For12] Otto Forster. *Analysis 3: Maß-und Integrationstheorie, Integralsätze im $\mathbb{R}^n$ und Anwendungen*, volume 3. Springer-Verlag, 2012.

[For17] Otto Forster. *Analysis 2: Differentialrechnung im $\mathbb{R}^n$, gewöhnliche Differentialgleichungen*. Springer-Verlag, 2017.

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JS04] Florian Jarre and Josef Stoer. *Optimierung*. Springer, 2004.

[KA16] Leonid Vitalevich Kantorovich and Gleb Pavlovich Akilov. *Functional Analysis*. Elsevier, 2016.

[KL20] Patrick Kidger and Terry Lyons. Universal Approximation with Deep Narrow Networks. In *Conference on learning theory*, pages 2306–2327. PMLR, 2020.

[LL01] Elliott H Lieb and Michael Loss. *Analysis*, volume 14. American Mathematical Soc., 2001.

[LLS22] Qianxiao Li, Ting Lin, and Zuowei Shen. Deep Learning via Dynamical Systems: An Approximation Perspective. *Journal of the European Mathematical Society*, 2022.

[LPW+17] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The Expressive Power of Neural Networks: A View from the Width. *Advances in neural information processing systems*, 30, 2017.

[TTI+20] Takeshi Teshima, Koichi Tojo, Masahiro Ikeda, Isao Ishikawa, and Kenta Oono. Universal Approximation Property of Neural Ordinary Differential Equations. *arXiv preprint arXiv:2012.02414*, 2020.

[Wer06] Dirk Werner. *Funktionalanalysis*. Springer-Verlag, 2006.

Appendix

A Supplement Material to Section 2

A.1 Sobolev Norm greater than Evaluation

Lemma. *Let $\Theta \in \mathcal{W}^\infty$. Then, for any $t \in [0,1]$, it holds that*

$$|\Theta(t)| \leq ||\Theta||_{H^1}.$$

Proof of Lem.2.4. The function Θ is continuous by Sobolev embeddings (see [LL01] for details). Since $[0,1]$ is compact, we thus find $t^* \in [0,1]$, such that for any $t \in [0,1]$ we have $|\Theta(t^*)| \leq |\Theta(t)|$. Furthermore, we can apply the fundamental theorem of calculus for H^1 and the triangle-inequality to obtain for any $t \in [0,1]$

$$\begin{aligned} |\Theta(t)| &= \left|\Theta(t^*) + \int_{t^*}^{t} \dot{\Theta}(s)ds\right| \\ &\leq |\Theta(t^*)| + \int_{t^*}^{t} |\dot{\Theta}(s)|ds. \end{aligned}$$

Augmenting the integral boundaries and using the fact that $\Theta(t^*)$ is a minimum we get

$$\begin{aligned} |\Theta(t)| &\leq \int_0^1 |\Theta(t^*)|ds + \int_0^1 |\dot{\Theta}(s)|ds \\ &\leq \int_0^1 |\Theta(s)|ds + \int_0^1 |\dot{\Theta}(s)|ds. \end{aligned}$$

Applying Hölder's inequality on both integrals on the right side of the last inequality we get

$$\begin{aligned} |\Theta(t)| &\leq ||\Theta||_{2,[0,1]} + ||\dot{\Theta}||_{2,[0,1]} \\ &= ||\Theta||_{H^1}. \end{aligned}$$

□

A.2 Properties of ODE-Flows

Theorem. *Let $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$, $\Theta \in \mathcal{W}^\infty$ and $x, x' \in \mathbb{R}^d$. Then, the following statements hold:*

1. $Z_0(x, \Theta) = x$
2. $Z_\bullet(x, \Theta) \in \mathcal{C}^1([0,1]; \mathbb{R}^d)$ *and* $l(Z_\bullet(x, \Theta)) = \int_0^1 |f(Z_t(x, \Theta), \Theta(t))| dt$
3. *Trajectories do not intersect: If $x \neq x'$ then for any $t \in [0,1]$ we has $Z_t(x, \Theta) \neq Z_t(x', \Theta)$.*
4. *Lipschitz-continuity in initial data: There exists a constant $C_{f,\Theta} > 0$, such that for any $t \in [0,1]$ on has*
$$|Z_t(x, \Theta) - Z_t(x', \Theta)| \leq C_{f,\Theta} \cdot |x - x'|.$$
5. *Homeomorphism in initial data: For any $t \in [0,1]$ the map $Z_t(\cdot, \Theta)$ is a homeomorphism on $\mathbb{R}^d$.*

Proof of Thm.2.8. Properties *1.* and *2.* are straight-forward. Number *3.* is well-known and can be found in [Aul04]. Number *5.* is [DDT19, Proposition 3]. The Lipschitz-estimate we prove ourselves. Let $t \in [0,1]$ and $x, x' \in \mathbb{R}^d$. Put

$$u(t) := |Z_t(x, \Theta) - Z_t(x', \Theta)|.$$

Using the fundamental theorem of calculus, the triangle inequality and the fact that $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$, we get

$$\begin{aligned} u(t) &\leq |x - x'| + \int_0^t |f(Z_s(x, \Theta), \Theta(s)) - f(Z_s(x', \Theta), \Theta(s))| \, ds \\ &\leq |x - x'| + \int_0^t \gamma_f(|\Theta(s)|) \cdot u(s) ds. \end{aligned}$$

Using Lem.2.4 yields

$$u(t) \leq |x - x'| + \gamma_f(||\Theta||_{H^1}) \cdot \int_0^t u(s) ds.$$

Lastly, with Grönwall's inequality we can conclude

$$u(t) \leq C_{f,\Theta} \cdot |x - x'|,$$

where $C_{f,\Theta} := e^{\gamma_f(||\Theta||_{H^1})}$. This finishes the proof. □

A.3 Well-definedness of ResNets

Definition/Lemma. *Let $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ and $L \in \mathbb{N}$. Define the* discrete flow *as the map*

$$\begin{aligned} z_\bullet[\cdot, \star] : \{0, \dots, L\} \times \mathbb{R}^d \times \mathcal{W}^L &\to \mathbb{R}^d \\ (l, x, \theta^L) &\mapsto z_l[x, \theta^L], \end{aligned}$$

via the recursion

$$\begin{aligned} z_{l+1}[x, \theta^L] &= z_l[x, \theta^L] + f(z_l[x, \theta^L], \theta^L_l), \quad l = 0, \dots, L-1 \\ z_0[x, \theta^L] &:= x. \end{aligned}$$

This function is well-defined and continuous for all $l = 0, \dots, L$.

Proof of Def./Lem.2.12: We need to show well-definedness and continuity, which we will do via induction. So firstly, let $L = 1$. Then, for any $x \in \mathbb{R}^d$ and $\theta^1 \in \mathcal{W}$ by definition we have

$$z_1[x, \theta^1] = x + f(x, \theta^1),$$

which is clearly well-defined and continuous in the first and second argument, since $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$. Now, suppose the claim holds for $L \in \mathbb{N}$ arbitrary but fixed. Then, by definition

$$z_{L+1}[x, \theta^{L+1}] = z_l[x, \theta^{L+1}] + f(z_l[x, \theta^{L+1}], \theta^{L+1}_L)$$

for any $x \in \mathbb{R}^d$ and $\theta^{L+1} \in \mathcal{W}^{L+1}$. This is is again continuous and well-defined in both arguments by the induction hypothesis and the given properties of f. □

B Supplement Material to Section 3

B.1 Extension of Original Universal Approximation Theorem

Corollary. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ be continuous with*

$$\lim_{s\to+\infty} \sigma(s) = 1 \quad \textit{and} \quad \lim_{s\to-\infty} \sigma(s) = 0.^{14} \tag{.2}$$

Moreover, let $K \subset \mathbb{R}^d$ be a compact subset and $h : K \to \mathbb{R}$ a continuous function. Then, for every $\epsilon > 0$ there exists an $N \in \mathbb{N}$ and $s_n, d_n \in \mathbb{R}$, $c_n \in \mathbb{R}^d$, for every $n = 1, \dots, N$, such that the function

$$v : K \to \mathbb{R}, \quad x \mapsto \sum_{n=1}^{N} s_n \sigma(c_n \cdot x + d_n)$$

satisfies $||h - v||_{\infty,K} < \epsilon$.

Proof of Cor.3.2. Since h is continuous, we know that every component $h_i : K \to \mathbb{R}$ is continuous for $i = 1, \dots, m$. By Thm.3.1 we find for every such i a positive integer $N_i \in \mathbb{N}$ and $s_{n,i}, d_{n,i} \in \mathbb{R}$, $c_{n,i} \in \mathbb{R}^d$, for every $n = 1, \dots, N_i$, such that the funciton defined by

$$v_i : K \to \mathbb{R}, \quad x \mapsto \sum_{n=1}^{N_i} s_{l,i} \sigma(c_{n,i} \cdot x + d_{n,i})$$

satisfies $||h_i - v_i||_{\infty,K} < \epsilon$. Note that we can define $N := \max_{i=1,\dots,m} N_i$ and consider without loss of generality functions v_i, where we sum from $n = 1$ up to $n = N$, by simply adding zeroes to the sum given in the definition of v_i above. Next define the vectors

$$s_n := (s_{n,1}, \dots, s_{n,m}), \quad d_n := (d_{n,1}, \dots, d_{n,m}) \in \mathbb{R}^m$$

and the matrices $C_n \in \mathbb{R}^{m\times d}$, that have the vectors $c_{n,1}, \dots, c_{n,m} \in \mathbb{R}^d$ as rows. With this we put the function $v = (v_1, \dots, v_m) : K \to \mathbb{R}^m$ to be

$$v(x) = \sum_{n=1}^{N} s_n \odot \sigma(C_n \cdot x + d_n)$$

for all $x \in K$. It is now left to show, that v approximates h ϵ-well w.r.t. the sup-norm. This follows swiftly from the fact that all norms on $\mathbb{R}^m$ are equivalent and the definition of the sup-norm. Indeed, by the equivalence of norms in finite dimensional vector spaces, there exists a constant $c > 0$, such that for any $x \in K$ we have

$$|h(x) - v(x)| \le c \cdot \max_{i=1,\dots,m} |h_i(x) - v_i(x)| \le c \cdot \max_{i=1,\dots,m} ||h_i - v_i||_{\infty,K}.$$

Hence,

$$||h - v||_{\infty,K} \le c \cdot \max_{i=1,\dots,m} ||h_i - v_i||_{\infty,K} < c \cdot \epsilon,$$

which becomes small, when ϵ gets small. This finishes the proof. □

[14] In the literature, this property is often called 'being a *sigmoidal*'.

B.2 Technical Lemmas

Again, all the proofs for this part of the Appendix are taken from [AK20].

Lemma. *Let $m \leq d$, $s, d \in \mathbb{R}^m$ and $C \in \mathbb{R}^{m \times d}$ with no zero rows. Then, there exist $\tilde{s}_l, \tilde{d}_l \in \mathbb{R}^m$ and $\tilde{C}_l \in \mathbb{R}^{m \times d}$, $l = 1, \ldots, m$, such that*

$$s \odot \sigma(Cx + d) = \sum_{l=1}^{m} \tilde{s}_l \odot \sigma(\tilde{C}_l x + \tilde{d}_l),$$

for any $x \in \mathbb{R}^d$, with $\operatorname{rank}(\tilde{C}_l) = m$, for $l = 1, \ldots, m$. If $m = d$ we can choose the matrices $\tilde{C}_l$, such that $\det(\tilde{C}_l) > 0$.

Proof of Lem.3.16. Let $m \leq d$. For any $1 \leq l \leq m$ we can find a matrix $\tilde{C}_l \in \mathbb{R}^{m \times d}$ with full rank, such that the l-th row of $\tilde{C}_l$ and C concur. We can find such a matrix by putting the l-th row to be the l-th row of C. By a well-known result form Linear Algebra, we can then choose $m - 1$ row vectors in $\mathbb{R}^d$, such that the collection of rows in $\tilde{C}_l$ are linearly independent and hence, $\operatorname{rank}(\tilde{C}_l) = m$. Moreover, we define the k-th component of $\tilde{s}_l$ and $\tilde{d}_l$, $k = 1, \ldots, d$ as

$$(\tilde{s}_l)_k = \begin{cases} s_k, & l = k \\ 0, & l \neq k \end{cases}, \qquad (\tilde{d}_l)_k = \begin{cases} d_k, & l = k \\ 0, & l \neq k \end{cases}.$$

Then, by construction and the fact that σ acts componentwise, we get for the k-th component of the right hand side of (3.8)

$$\begin{aligned} \left(\sum_{l=1}^{m} \tilde{s}_l \odot \sigma(\tilde{C}_l x + \tilde{d}_l) \right)_k &= \sum_{l=1}^{m} (\tilde{s}_l \odot \sigma(\tilde{C}_l x + \tilde{d}_l))_k \\ &= \sum_{l=1}^{m} (\tilde{s}_l)_k \cdot (\sigma(\tilde{C}_l x + \tilde{d}_l))_k \\ &= \sum_{l=1}^{m} (\tilde{s}_l)_k \cdot \sigma((\tilde{C}_l x)_k + (\tilde{d}_l)_k) \\ &= s_k \cdot \sigma((Cx)_k + d_k) \\ &= (s \odot \sigma(Cx + d))_k. \end{aligned}$$

Thus, the equality is shown. Additionally, if $m = d$, we know, since $\operatorname{rank}(\tilde{C}_l) = m$, that $\det(\tilde{C}_l) \neq 0$. Assuming $\det(\tilde{C}_l) < 0$, we can, without loss of generality, multiply a row of $\tilde{C}_l$ (other than the l-th one) by -1 to get $\det(\tilde{C}_l) > 0$. This finishes the proof. □

Lemma. *Let $m \leq d$, $L \in \mathbb{N}$, and $s_l, d_l \in \mathbb{R}^m$ and $C_l \in \mathbb{R}^{m\times d}$ for $l = 1, \dots, L$ (each C_l having no zero rows). Then, there exists an $\tilde{L} \in \mathbb{N}$ and $\tilde{s}_k, \tilde{d}_k \in \mathbb{R}^m$, $\tilde{C}_k \in \mathbb{R}^{m\times d}$ for $k = 1, \dots, \tilde{L}$, such that*

$$\sum_{l=1}^{L} s_l \odot \sigma(C_l x + d_l) = \sum_{k=1}^{\tilde{L}} \tilde{s}_k \odot \sigma(\tilde{C}_k x + \tilde{d}_k),$$

for any $x \in \mathbb{R}^d$, with $\operatorname{rank}(\tilde{C}_k) = m$, for $k = 0, \dots, \tilde{L} - 1$. If $m = d$ we can choose the matrices $\tilde{C}_k$, such that $\det(\tilde{C}_k) > 0$.

Proof of Lem.3.17. This follows immediately by applying Lem.3.16 to the summands on the right hand side of (3.9). □

Lemma. *Let $m \leq d$. Let $C \in \mathbb{R}^{m\times d}$ with $\operatorname{rank}(C) = m$. If $m = d$ assume additionally $\det(C) > 0$. Then, there exists a $P \in \mathbb{R}^{d\times d}$, such that*

$$C = \mathbb{M}P \quad \textit{and} \quad \det(P) > 0.$$

Proof of Lem.3.18. For the case $m = d$ with the restriction $\det(C) > 0$, $\mathbb{M}$ is by definition the identity matrix and after choosing $P = C$ there is nothing to prove. So let $m < d$. Since C has full rank, with a similar argument as in the proof of Lem.3.16, there exist $d - m$ row vectors in $\mathbb{R}^d$, such that if we put $P \in \mathbb{R}^{n\times n}$ to be the matrix, whose first m rows are the ones from C and the last $d - m$ rows are the chosen row vectors, we get $\det(P) \neq 0$. Again, without loss of generality, we may choose $\det(P) > 0$. Furthermore, multiplying with $\mathbb{M}$ corresponds to cancelling out everything of a matrix besides the first m rows. this gives $\mathbb{M}P = C$, which is the claim. □

Lemma. *Let $p \in [0, \infty[$ and $L \in \mathbb{N}$. For $P_l \in \mathbb{R}^{d\times d}$, $l = 1, \dots, L$ with $\det(P_l) > 0$ and $0 = t_0 < t_1 < \cdots < t_L = 1$ define the piecewise constant function*

$$P : [0, 1] \to \mathbb{R}^{d\times d}, \quad t \mapsto P(t) := P_L \mathbb{1}_{\{1\}}(t) + \sum_{l=1}^{L} P_l \mathbb{1}_{[t_{l-1}, t_l[}(t).$$

Then, there exists a $c > 0$, such that for any $\eta > 0$, there exists a $P_\eta \in \mathcal{C}^\infty([0, 1]; \mathbb{R}^{d\times d})$ with

$$||P_\eta - P||_{p,[0,1]} < \eta, \quad \det(P_\eta(t)) > 0, \quad |P_\eta(t)| \leq c$$

for all $t \in [0, 1]$.

Proof of Lem.3.19. Let $\eta > 0$. We define the space of invertible $d \times d$-matrices with strictly positive determinant as $\mathrm{GL}_d^+(\mathbb{R})$. Note that this space is path-connected (see [Bak03, Chapter 9, p. 239]), so for any $l = 1, \dots, L$ we can find a continuous map $Q^{(l)} : [0,1] \to \mathrm{GL}_d^+(\mathbb{R})$, such that

$$Q^{(l)}(0) = P^{(l)} \quad \text{and} \quad Q^{(l)}(1) = P^{(l+1)}.$$

We now connect all these paths and define a new path on $\mathbb{R}$. For $\delta > 0$ put

$$Q^\delta(t) := \begin{cases} P^{(1)}, & -\infty < t < t_1, \\ Q^{(l)}\left(\frac{t-t_l}{\delta}\right), & t_l \le t < t_l + \delta, \qquad \text{for } l = 1, \dots, L-1 \\ P^{(l)}, & t_{l-1} + \delta \le t < t_l, \quad \text{for } l = 2, \dots, L-2 \\ P^{(L)}, & t_{L-1} + \delta \le t < \infty. \end{cases}$$

Note that $Q^\delta \in \mathcal{C}(\mathbb{R}, \mathrm{GL}_d^+(\mathbb{R}))$. For $\epsilon > 0$ let $\phi_\epsilon : \mathbb{R} \to \mathrm{GL}_d^+(\mathbb{R})$ be a standard mollifier and put

$$P_\epsilon := \phi_\epsilon * Q^\delta.$$

Notice that Q^δ is continuous on $[0,1]$ and on such compact sets the mollifications P_ϵ converge uniformly. Since Q_δ maps to $\mathrm{GL}_d^+(\mathbb{R})$, choosing ϵ small enough gives $\det(P_\epsilon(t)) > 0$ for any $t \in [0,1]$. By the triangle inequality we get

$$||P_\epsilon - P||_{p,[0,1]} \le ||P_\epsilon - Q^\delta||_{p,[0,1]} + ||Q^\delta - P||_{p,[0,1]},$$

and the rest of the proof now consists of showing, that the right hand side of this inequality gets small with δ and ϵ. By properties of the mollification we choose ϵ small enough, such that

$$||P_\epsilon - Q^\delta||_{p,[0,1]} < \frac{\eta}{2},$$

For the second term consider the following equations, where we used the definition of Q^δ and the transformation formula:

$$\begin{aligned} ||Q^\delta - P||_{p,[0,1]}^p &= \int_0^1 |Q^\delta(t) - P(t)|^p dt \\ &= \sum_{l=1}^{L-1} \int_{t_l}^{t_l+\delta} \left| Q^{(l)}\left(\frac{t-t_l}{\delta}\right) - P^{(l+1)} \right|^p dt \\ &= \delta \sum_{l=1}^{L-1} \int_0^1 |Q^{(l)}(s) - P^{(l+1)}|^p ds. \end{aligned}$$

Since $|Q^{(l)}(s) - P^{(l+1)}|$ is bounded (uniformly in s and l), this quantity also vanishes as $\delta \to 0$. So choosing δ and small enough and putting $P_\eta := P_\epsilon$ gives the claim that

$$||P_\eta - P||_{p,[0,1]} < \eta.$$

Note that P_η is continuous, since it is smooth, so, by compactness of $[0,1]$, there exists a $C > 0$, such that $|P_\eta(t)| \le C$ for any $t \in [0,1]$. Thus, the proof is finished. □

B.3 Approximation Speed of Augmented Neural ODE

To prove the claims made in Rem.3.20 about the approximation speed/ complexity bound of the model we reduce the claim without loss of generality to the case of approximating the function $h : \mathbb{R} \to \mathbb{R}$, $t \mapsto \mathbb{1}_{[0,1]}(t)$ with mollifications w.r.t. $||\cdot||_{1,[0,1]}$ (see (3.11)). In general the approximation speed of such mollifications can be arbitrarily slow/fast. However, the functions s and q chosen in the proof of Thm.3.15 are sums of indicator functions, so, by the triangle inequality, this choice of h is reasonable. For the following proof we let $(\phi_\eta)_{\eta>0}$ be a sequence of standard mollifiers. We will show that for $h_\eta := h * \phi_\eta$ there exists a $c > 0$, such that

$$||h_\eta - h|| \leq c \cdot \eta.$$

Proof of the Approximation Bound. Note that $h(t) = \mathbb{1}_{[0,\infty[}(t) - \mathbb{1}_{[0,\infty[}(t-1)$ for all $t \in \mathbb{R}$. Put $r := \mathbb{1}_{[0,\infty[}$. Using the transformation formula, we then have

$$\begin{aligned} r_\eta(t) &= (\phi_\eta * r)(t) \\ &= \frac{1}{\eta} \int_{\mathbb{R}} r(s)\phi\left(\frac{t-s}{\eta}\right) ds \\ &= \int_0^\infty \phi\left(\frac{t}{\eta} - s\right) ds \\ &= r_1\left(\frac{t}{\eta}\right). \end{aligned}$$

Notice that for any $\alpha > 0$ we have $r(t) = r(\alpha t)$ for any $t \in \mathbb{R}$. For $\alpha = \frac{1}{\eta}$ we thus get using the transformation formula again

$$\begin{aligned} ||r_\eta - r||_1 &= \int_{\mathbb{R}} |r_\eta(t) - r(t)| dt \\ &= \int_{\mathbb{R}} \left| r_1\left(\frac{t}{\eta}\right) - r\left(\frac{t}{\eta}\right) \right| dt \\ &= \eta \cdot ||r_1 - r||_1. \end{aligned}$$

Observe that $||r_1 - r||_1$ is finite, since

$$r_1(t) - r(t) = \left(\int_{-\infty}^t \phi(\tilde{s}) d\tilde{s} \right) - \mathbb{1}_{[0,\infty[}(t),$$

which is bounded and compactly supported. Moreover, $||r_1 - r||_1$ independent of η. Hence, we found $c > 0$, such that

$$\begin{aligned} ||r_\eta - r||_{1,[0,1]} &\leq ||r_\eta - r||_1 \\ &= c \cdot \eta. \end{aligned}$$

Using $h = r - r(\cdot - 1)$ and the linearity of the convolution yields the same estimate for h. $\square$

C Supplement Material to Section 4

C.1 Definition of Fréchet differentiability

The following definition is taken from [Wer06].

Definition. *Let $(X, ||\cdot||_X), (Y, ||\cdot||_Y)$ be two normed vector spaces and $U \subset X$ be open. A function $F : U \to Y$ is called* Fréchet differentiable *at $x \in U$, if there exists a bounded linear operator $P : X \to Y$, such that*

$$\lim_{||h||_X \to 0} \frac{||F(x+h) - F(x) - Ph||_Y}{||h||_X} = 0.$$

Put differently this means that

$$||F(x+h) - F(x) - Ph||_Y \in o(||h||_X).$$

We also write $dF_x := P$.

C.2 Continuity w.r.t. Weight Function

Lemma. *Let $f \in \mathcal{F}(\mathbb{R}^d, \mathcal{W})$ be continuously differentiable (in both variables) with bounded derivative. Let $\Theta, \Delta \in \mathcal{W}^\infty$ and $s \in]0,1[$. Then, there exists $c > 0$, such that*

$$|(Z_s(x, \Theta + \Delta) - Z_s(x, \Theta), \Delta(s))| \leq c \cdot ||\Delta||_{H^1}.$$

Proof. We use Lem.2.4 and equivalence of norms on finite dimensional vector spaces, to find $c' > 0$, such that

$$\begin{aligned} |(Z_s(x, \Theta + \Delta) - Z_s(x, \Theta), \Delta(s))| &\leq c' \cdot (|Z_s(x, \Theta + \Delta) - Z_s(x, \Theta)| + |\Delta(s)|) \\ &\leq c' \cdot |Z_s(x, \Theta + \Delta) - Z_s(x, \Theta)| + c'||\Delta||_{H^1}. \end{aligned}$$

Hence, proving the statement reduces itself to proving continuity of $Z_s(x, \star)$. Put

$$w(\Delta, s) := |Z_s(x, \Theta + \Delta) - Z_s(x, \Theta)|.$$

Using the fundamental theorem of calculus, the triangle inequality, the mean value inequality and the fact that the derivative of f is bounded by $M > 0$, we get

$$\begin{aligned} w(\Delta, s) &\leq \int_0^s |f(Z_r(x, \Theta + \Delta), \Theta(r) + \Delta(r)) - f(Z_r(x, \Theta), \Theta(r))| dr \\ &\leq M \cdot \int_0^s |(Z_r(x, \Theta + \Delta) - Z_r(x, \Theta), \Delta(r))| dr. \end{aligned}$$

Again, norms are equivalent in finite dimensions, and thus,, after applying Lem.2.4, it follows that

$$\begin{aligned} w(\Delta, s) &\leq c'M \cdot \int_0^s |\Delta(r)| dr + c'M \cdot \int_0^s w(\Delta, r) dr \\ &\leq c'M \cdot ||\Delta||_{H^1} + c'M \cdot \int_0^s w(\Delta, r) dr. \end{aligned}$$

Finally we apply Grönwall's inequality to obtain

$$w(\Delta, s) \leq c' M e^{c' M} \cdot ||\Delta||_{H^1}.$$

Combining everything yields

$$|(Z_s(x, \Theta + \Delta) - Z_s(x, \Theta), \Delta(s))| \leq \left((c')^2 M e^{c' M} + c' \right) \cdot ||\Delta||_{H^1} =: c \cdot ||\Delta||_{H^1},$$

which is what we had to prove. □

D Supplement Material to Section 5

D.1 Lem.2.4 fails for Lebesgue Norm

We give an example of a function $q \in L^2([0,1],\mathbb{R})$, such that there exists a set $S \subset [0,1]$ with non-zero Lebesgue measure, so that for any $t \in S$ we have

$$||q||_{L^2} < q(t).$$

For $n \in \mathbb{N}$ we put

$$q_n(t) := \begin{cases} 4n, & t \in [0, \frac{1}{4n}[\\ 1, & t \in [\frac{1}{4n}, \frac{3}{4n}[\\ \frac{4n}{4n-3} \cdot (t - \frac{1}{n}), & t \in [\frac{3}{4n}, \frac{1}{n}]. \end{cases}$$

An easy calculation shows

$$||q_n||_{2,[0,1]} = \sqrt{\frac{2}{3n}}.$$

Choose $m \in \mathbb{N}$ large enough, such that $||q_m||_{2,[0,1]} < 1$. Furthermore, for $S := [\frac{1}{4m}, \frac{3}{4m}[$, which has non-zero Lebesgue measure, we have $q_m\big|_S = 1$. Hence,

$$||q_m||_{2,[0,1]} < q_m(t),$$

for any $t \in S$ and we have found a counterexample.