ON THE SPACE-TIME EXPRESSIVITY OF RESNETS

Johannes Müller
Max Planck Insitute for Mathematics in the Sciences
jmueller@mis.mpg.de

ABSTRACT

Residual networks (ResNets) are a deep learning architecture that substantially improved the state of the art performance in certain supervised learning tasks. Since then, they have received continuously growing attention. ResNets have a recursive structure $x_{k+1} = x_k + R_k(x_k)$ where R_k is a neural network called a residual block. This structure can be seen as the Euler discretisation of an associated ordinary differential equation (ODE) which is called a neural ODE. Recently, ResNets were proposed as the space-time approximation of ODEs which are not of this neural type. To elaborate this connection we show that by increasing the number of residual blocks as well as their expressivity the solution of an arbitrary ODE can be approximated in space and time simultaneously by deep ReLU ResNets. Further, we derive estimates on the complexity of the residual blocks required to obtain a prescribed accuracy under certain regularity assumptions.

1 INTRODUCTION

Various neural network based methods have been proposed for the numerical analysis of partial differential equations (PDEs) (see Lee and Kang, 1990; Dissanayake and Phan-Thien, 1994; Takeuchi and Kosugi, 1994; Lagaris et al., 1998) as well as for ordinary differential equations (ODEs) (see Meade Jr and Fernandez, 1994a;b; Lagaris et al., 1998; Breen et al., 2019). In subsequent years those methods where improved and extended to a variety of settings and we refer to Yadav et al. (2015) for an overview of neural network based methods for ODEs. Recently, deep networks have successfully been applied to the numerical simulation of stationary and non stationary PDEs by E and Yu (2018) and E et al. (2017); Han et al. (2018) respectively; a list of further improvement of those methods can be found in Grohs et al. (2019). The promising empirical performance of those approaches raised interest in theoretical guarantees and led to a number of error estimates (see Jentzen et al., 2018; Han and Long, 2018; Grohs et al., 2018; Elbrächter et al., 2018; Berner et al., 2018; Reisinger and Zhang, 2019; Kutyniok et al., 2019). In particular it can be shown that neural networks are capable of approximating the solutions of a number of PDEs without suffering from the curse of dimensionality. However, it should be noted that those works only provide estimates for the spatial error at a fixed time rather than the approximation error in space and time simultaneously.

Compared to the case of PDEs the analysis of the approximation error for ODEs is less complete. Although a priori and a posteriori error estimates are present in the literature (see Filici, 2008; 2010, respectively) they only consider the solution for a single initial value rather than the full space-time solution

$$x(0, y) = y, \quad \partial_t x(t, y) = f(t, x(t, y)) \quad \text{for all } t, y \tag{1}$$

to the right hand side f. Recently, Grohs et al. (2019) established an approximation result in space-time and showed that Euler discretisations of a certain class of neural ODEs can be approximated by neural networks with error decreasing exponentially in the complexity of the networks. Those are the first space-time error estimates in the study of neural network based methods for either PDEs or ODEs. Yet, in order to obtatin space-time error estimates for the solution of an ODE one has to bound the approximation error of the class of Euler discretisations to the solution of this ODE. Such estimates are implied by our main result Theorem 3 concerning the approximation of space-time solutions with residual networks. However, there is a further motivation in the study of the approximation capabilities of residual networks that we will present now.

Residual networks and dynamical systems

Residual networks (ResNets) make use of skip connections which were introduced to overcome difficulties in the training of deep neural networks in supervised learning tasks. Rather than using the iterative scheme $x_{l+1} \coloneqq \rho(A_l x_l + b_l)$ like a traditional feedforward network, ResNets copy the input x_l to some subsequent layer, in the easiest case to the following layer which leads to

$$x_{l+1} \coloneqq x_l + \rho(A_l x_l + b_l) \quad \text{for } l = 0, \dots, L-1. \tag{2}$$

Obviously, this is only well defined if the dimensions of all states x_l agree. It was shown in He et al. (2016) that ResNets are superior to traditional feedforward neural networks in some image classification tasks. It has been pointed out in Haber et al. (2018) that the recursive structure (2) can be interpreted as the explicit Euler discretisation of the ordinary differential equation

$$\partial_t x(t) = \rho\big(A(t)x(t) + b(t)\big). \tag{3}$$

Building on this observation Haber and Ruthotto (2017) transferred the knowledge about the stability of ODEs to the stability of forward propagation in ResNets and Lu et al. (2017) introduced neural networks corresponding to other numerical schemes for ODEs like implicit Euler or Runge-Kutta schemes. Further, Chen et al. (2018) replaced ResNets by ODEs in supervised learning tasks and achieved state of the art performance with fewer parameters. A rigorous justification for this approach using the notion of Γ-convergence was established in Thorpe and van Gennip (2018). Lately, ResNets have been proposed in Rousseau et al. (2019) as an approximation of space-time solutions of a much more general class of ODEs than (3) which always admits non decreasing solutions. Further, this was applied to the problem of diffeomorphic image registration which can be interpreted as a controlled ODE problem.

The expressivity of ResNets was studied in different ways including the following. It was shown by Lin and Jegelka (2018) that ResNets are able to approximate arbitrary L^p-functions and Cuchiero et al. (2019) showed that ResNets can take prescribed values on arbitrary point sets. Both works consider the input-output mapping $x_0 \mapsto x_L$ induced by a ResNet. Similarly, Dupont et al. (2019) and Zhang et al. (2019) studied the approximation capabilities of neural ODEs at final time. Although many works perceive ResNets as discrete dynamical systems (see E, 2017; Liu and Markowich, 2019, and subsequent work) an analysis of the expressivity of their dynamics is still absent.

Contributions

We study the expressivity of the dynamics of ResNets and show that ResNets can approximate solutions of arbitrary ODEs in space-time. This includes the solution of the control problem ResNets where proposed for in Rousseau et al. (2019). More precisely, we make the following contributions:

1. *Universality*: ResNets can approximate solutions of arbitrary ODEs uniformly in space and time simultaneously (see Theorem 2).
2. *Complexity bounds*: Assume that the right hand side f is Lipschitz continuous. Then the solution to this ODE can be approximated with (local) error $\mathcal{O}(n^{-1})$ through ResNets with n residual blocks which have $\mathcal{O}(r_n^d n^d)$ neurons; here, $(r_n)_{n\in\mathbb{N}} \subseteq (0,\infty)$ is an arbitrary sequence diverging to $+\infty$ and d is the dimension of the ODE (see Theorem 3).

2 Definitions and notation

Let for the remainder d, m, L be natural numbers. Further, we consider tupels

$$\theta = ((A_1, b_1), \dots, (A_L, b_L))$$

of matrix-vector pairs where $A_l \in \mathbb{R}^{N_l \times N_{l-1}}$ and $b_l \in \mathbb{R}^{N_l}$ and $N_0 = d, N_L = m$. Every matrix vector pair (A_l, b_l) induces an affine linear transformation that we denote by $T_l \colon \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}$. The *neural network with parameters* θ and with respect to some *activation function* $\rho \colon \mathbb{R} \to \mathbb{R}$ is the function

$$R = R_\theta \colon \mathbb{R}^d \to \mathbb{R}^m, \quad x \mapsto T_L(\rho(T_{L-1}(\rho(\cdots \rho(T_1(x)))))),$$

where ρ is applied componentwise. We call d the *input* and m the *output dimension*, L the *depth* and $N(\theta) \coloneqq \sum_{l=0}^{L} N_l$ the *number of neurons* of the network. If we have $f = R_\theta$ for some θ we say that the function f is *expressed* by the neural network.

In the following we restrict ourselves to the case of a specific activation function which is not only commonly used in practice (see Ramachandran et al., 2017) but also exhibits nice theoretical properties (see Arora et al., 2016; Petersen et al., 2018). The *rectified linear unit* or *ReLU activation function* is defined via $x \mapsto \max\{0, x\}$ and we call networks with this activation *ReLU networks*.

Finally, we introduce the notion of residual networks. In order to interpret ResNets as functions in space-time we define them to be Euler discretisation of a certain class of ODEs which are linearly interpolated in time. It is important to note that this might differ from other definitions of residual networks present in the literature.

Definition 1 (Residual network)**.** Let $\theta = (\theta_1, \dots, \theta_n)$ be a tupel of parameters of neural networks with input and output dimension d. Let $R_1, \dots, R_n \colon \mathbb{R}^d \to \mathbb{R}^d$ denote the neural networks with parameters $\theta_1, \dots, \theta_n$ and some activation ρ. We refer to those networks as *residual blocks*. The *residual network* or *ResNet* $x^n \colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ with parameters $\theta = (\theta_1, \dots, \theta_n)$ and with respect to the activation function ρ is defined via

$$x^n(0, y) \coloneqq y, \quad x^n(t_{k+1}, y) \coloneqq x^n(t_k, y) + n^{-1} \cdot R_{k+1}(x^n(t_k, y))$$

for $k = 0, \dots, n-1$ and linearly in between. In the remainder, we will only consider ResNets with respect to the ReLU activation function and call those *ReLU ResNets*.

3 Presentation of the main results

Now we have introduced enough notation to state our main results precisely.

Theorem 2 (Space-time approximation with ResNets)**.** *Let $d \in \mathbb{N}$, $f \in L^1([0,1]; \mathcal{C}_b^{0,1}(\mathbb{R}^d; \mathbb{R}^d))$[1] and let x be the space-time solution[2] of the ODE with right hand side f. Then for every compact set $K \subseteq \mathbb{R}^d$ and $\varepsilon > 0$ there is a ReLU ResNet $\tilde{x}$ such that*

$$\|\tilde{x}(t, y) - x(t, y)\| \le \varepsilon \quad \textit{for all } t \in [0,1], y \in K.$$

The proof is based on the observation that f can be approximated by functions that are piecewise constant in time on the intervals $[k/n, (k+1)/n)$. By standard continuity results for the solution operator of ODEs the approximation also holds for the associated space-time solutions and thus one can without loss of generality assume that f is piecewise constant. However, if f is merely integrable in time it is not possible to bound the number of constant regions. Hence, one can not bound the number of the residual blocks that is required in order to achieve a prescribed approximation accuracy under no temporal regularity assumptions. Nevertheless, in the proof of the result the approximation in space and in time are clearly separated and in fact the constructed residual blocks share weights depending on the temporal regularity of the right hand side. Hence, the same arguments can be used to establish bounds on the complexity of the residual networks like the following.

Theorem 3 (Space-time approximation with complexity bounds)**.** *Let $d \in \mathbb{N}$, $(r_n)_{n\in\mathbb{N}} \subseteq (0, \infty)$ be a sequence diverging to $+\infty$ and let $f \colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ be a bounded and Lipschitz continuous function. Let $x \colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ be the space-time solution of the ODE with right hand side f. Then for every $n \in \mathbb{N}$ there is a ReLU ResNet x^n with parameters $\theta^n = (\theta_1^n, \dots, \theta_n^n)$ such that the following are satisfied:*

1. Approximation: *For every compact set $K \subseteq \mathbb{R}$ it holds*

$$\sup_{t\in[0,1], y\in K} \|x^n(t, y) - x(t, y)\| \in \mathcal{O}(n^{-1}).$$

2. Complexity bounds: *Every residual block θ_k^n has depth $\lceil \log_2((d+1)!) \rceil + 2$ and satisfies*

$$N(\theta_k^n) \in \mathcal{O}\left(r_n^d n^d\right).$$

Finally, all but $\mathcal{O}\left(r_n^d n^d\right)$ weights can be fixed.

[1] Up to a technical measurability property (Bochner measurability) this means that $f(t, \cdot)$ is bounded and Lipschitz continuous for almost all t and that the uniform norm and Lipschitz constants are integrable.

[2] see (1); we use the notion of weak solutions introduced in the appendix; if f is continuous this coincides with the classical notion of a solution; further, the ODE is globally well posed for this class of right hand sides.

The theory of nonlinear approximation (see DeVore et al., 1989) implies that cn^{d+1} parameters are required in order to approximate $d+1$ dimensional Lipschitz at speed $\mathcal{O}(n^{-1})$. However, the assumption that x solves a differential equation enables us to establish the upper bound of $\mathcal{O}\left(r_n^d n^d\right)$ parameters. Here r_n can approach $+\infty$ arbitrarily slow and hence can be chosen to lie below n^{d+1}. Further, we shall note that in similar fashion further complexity estimates can be obtained if the temporal and spatial regularities are different.

Outline of the proof

In a nutshell the proof of the approximation results presented above relies on a combination of a spacial approximation result for ReLU networks and a Grönwall argument. We quickly present the key arguments of Theorem 3 and postpone any rigorous calculations to the appendix.

The proof is based on a variant of the universal approximation results in Hanin (2017) and Yarotsky (2018) and we follow He et al. (2018) for the construction of piecewise linear interpolations. This method achieves optimal rates under the assumption of continuous weight assignment which are also optimal for bounded depth networks (see DeVore et al., 1989; Yarotsky, 2018). Although faster approximation rates for deep networks of bounded width are established in Yarotsky (2018) we use the following result as it allows a direct control of the uniform norm of the networks. However, our arguments can be generalised to other universal approximation results.

Proposition 4 (Universal approximation under Lipschitz condition)**.** *Let $d, m \in \mathbb{N}$ and $r > 0$ and let $f\colon \mathbb{R}^d \to \mathbb{R}^m$ be Lipschitz continuous. Then for every $\varepsilon > 0$ there is a ReLU network R_ε with parameters θ_ε that satisfies the following:*

1. Approximation: *It holds that* $\sup_{x\in[-r,r]^d} \|f(x) - R_\varepsilon(x)\| \le \varepsilon$.
2. Complexity bounds: *The network has depth* $\lceil \log_2((d+1)!)\rceil + 2$, $\mathcal{O}\left(r^d\varepsilon^{-d}\right)$ *many neurons and all but* $\mathcal{O}\left(r^d\varepsilon^{-d}\right)$ *weights can be fixed. Finally, if* $\|f\|$ *is bounded by* c *so is* $\|R_\varepsilon\|$.

Proof of Theorem 3. For every $n \in \mathbb{N}$ the previous proposition yields the existence of neural networks $R_1^n, \dots, R_n^n$ of asserted complexity that satisfy

$$\sup_{x\in[-r_n,r_n]^d} \left\|f(t_k, x) - R_{k+1}^n(x)\right\| \le n^{-1}.$$

Since f is bounded, let's say by $c > 0$[3], so are all realisations R_k^n independent of k and n. Hence, for any initial condition $y \in B_R$ in some ball the true solution $x(t, y)$ as well as the ResNet $x^n(t, y)$ arising from the networks $R_1^n, \dots, R_n^n$ remain in the bounded set B_{R+c}. However, on this bounded set the realisations R_k^n approximate the right hand side uniformly and thus every ResNet can be interpreted as an perturbed Euler discretisation of the ODE with right hand side f. Therefore, the residual network satisfies an integral equation for every fixed inital value y. An application of Grönwall's inequality yields that x^n does in fact converge towards x uniformly on $[0,1] \times B_R$ with approximation error in $\mathcal{O}(n^{-1})$. Since the ball B_R was arbitrary the general statement follows. □

4 Discussion and further research

We showed that residual networks are capable of approximating the solution of general ODEs in space-time. Further, under additional regularity assumptions we established bounds on the complexity of the residual blocks. The arguments presented above can directly be generalised to other classes of right hand sides f that allow a more effective spatial approximation through neural networks. This includes compositional functions or classes of (piecewise) smooth functions (see Mhaskar et al., 2016; Liang and Srikant, 2016; Petersen and Voigtlaender, 2018; Yarotsky, 2018; Shen et al., 2019; Montanelli and Yang, 2019).

For future research we propose to investigate whether the solutions of optimal control problems can be approximated efficiently by ResNets. Further, we believe it could be interesting to extend the results in Thorpe and van Gennip (2018); Avelin and Nyström (2019) where the correspondence of regularisers for ResNets and ODEs is investigated.

[3] By this we mean that the (Euclidean) norm is bounded by c.

ACKNOWLEDGMENTS

JM acknowledges support by the Evangelisches Studienwerk Villigst e.V. and the IMPRS MiS. Further, the authors want to thank Nihat Ay, Nicolas Charon, Philipp Harms, Jasper Hofmann and Guido Montúfar and Hsi-Wei Hsieh for valuable comments and discussions.

REFERENCES

W. Arendt, C. J. Batty, M. Hieber, and F. Neubrander. *Vector-valued Laplace transforms and Cauchy problems*, volume 96. Springer Science & Business Media, 2011.

R. Arora, A. Basu, P. Mianjy, and A. Mukherjee. Understanding Deep Neural Networks with Rectified Linear Units. *arXiv preprint arXiv:1611.01491*, 2016.

B. Avelin and K. Nyström. Neural ODEs as the Deep Limit of ResNets with constant weights. *arXiv preprint arXiv:1906.12183*, 2019.

J. Berner, P. Grohs, and A. Jentzen. Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations. *arXiv preprint arXiv:1809.03062*, 2018.

P. G. Breen, C. N. Foley, T. Boekholt, and S. P. Zwart. Newton vs the machine: solving the chaotic three-body problem using deep neural networks. *arXiv preprint arXiv:1910.07291*, 2019.

T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural Ordinary Differential Equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.

C. Cuchiero, M. Larsson, and J. Teichmann. Deep neural networks, generic universal interpolation, and controlled ODEs. *arXiv preprint arXiv:1908.07838*, 2019.

R. A. DeVore, R. Howard, and C. Micchelli. Optimal nonlinear approximation. *Manuscripta mathematica*, 63(4):469–478, 1989.

J. Diestel and J. Uhl. Vector Measures, 1977.

M. Dissanayake and N. Phan-Thien. Neural-network-based approximations for solving partial differential equations. *communications in Numerical Methods in Engineering*, 10(3):195–201, 1994.

E. Dupont, A. Doucet, and Y. W. Teh. Augmented Neural ODEs. *arXiv preprint arXiv:1904.01681*, 2019.

W. E. A Proposal on Machine Learning via Dynamical Systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

W. E and B. Yu. The Deep Ritz Method: A Deep Learning-Based Numerical Algorithm for Solving Variational Problems. *Communications in Mathematics and Statistics*, 6(1):1–12, 2018.

W. E, J. Han, and A. Jentzen. Deep Learning-Based Numerical Methods for High-Dimensional Parabolic Partial Differential Equations and Backward Stochastic Differential Equations. *Communications in Mathematics and Statistics*, 5(4):349–380, 2017.

D. Elbrächter, P. Grohs, A. Jentzen, and C. Schwab. DNN Expression Rate Analysis of High-dimensional PDEs: Application to Option Pricing. *arXiv preprint arXiv:1809.07669*, 2018.

C. Filici. On a Neural Approximator to ODEs. *IEEE transactions on neural networks*, 19(3):539–543, 2008.

C. Filici. Error estimation in the neural network solution of ordinary differential equations. *Neural Networks*, 23(5):614–617, 2010.

P. Grohs, F. Hornung, A. Jentzen, and P. Von Wurstemberger. A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations. *arXiv preprint arXiv:1809.02362*, 2018.

P. Grohs, F. Hornung, A. Jentzen, and P. Zimmermann. Space-time error estimates for deep neural network approximations for differential equations, 2019.

E. Haber and L. Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1): 014004, 2017.

E. Haber, L. Ruthotto, E. Holtham, and S.-H. Jun. Learning Across Scales—Multiscale Methods for Convolution Neural Networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

J. Han and J. Long. Convergence of the Deep BSDE Method for Coupled FBSDEs. *arXiv preprint arXiv:1811.01165*, 2018.

J. Han, A. Jentzen, and W. E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

B. Hanin. Universal Function Approximation by Deep Neural Nets with Bounded Width and ReLU Activations . *arXiv preprint arXiv:1708.02691*, 2017.

J. He, L. Li, J. Xu, and C. Zheng. ReLU Deep Neural Networks and Linear Finite Elements. *arXiv preprint arXiv:1807.03973*, 2018.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

A. Jentzen, D. Salimova, and T. Welti. A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of Kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients. *arXiv preprint arXiv:1809.07321*, 2018.

G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider. A Theoretical Analysis of Deep Neural Networks and Parametric PDEs. *arXiv preprint arXiv:1904.00377*, 2019.

I. E. Lagaris, A. Likas, and D. I. Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE transactions on neural networks*, 9(5):987–1000, 1998.

H. Lee and I. S. Kang. Neural algorithm for solving differential equations. *Journal of Computational Physics*, 91(1):110–131, 1990.

S. Liang and R. Srikant. Why Deep Neural Networks for Function Approximation? *arXiv preprint arXiv:1610.04161*, 2016.

H. Lin and S. Jegelka. ResNet with one-neuron hidden layers is a Universal Approximator. In *Advances in Neural Information Processing Systems*, pages 6169–6178, 2018.

H. Liu and P. Markowich. Selection dynamics for deep neural networks. *arXiv preprint arXiv:1905.09076*, 2019.

Y. Lu, A. Zhong, Q. Li, and B. Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. *arXiv preprint arXiv:1710.10121*, 2017.

A. J. Meade Jr and A. A. Fernandez. The numerical solution of linear ordinary differential equations by feedforward neural networks. *Mathematical and Computer Modelling*, 19(12):1–25, 1994a.

A. J. Meade Jr and A. A. Fernandez. Solution of nonlinear ordinary differential equations by feedforward neural networks. *Mathematical and Computer Modelling*, 20(9):19–44, 1994b.

H. Mhaskar, Q. Liao, and T. Poggio. Learning Functions: When Is Deep Better Than Shallow. *arXiv preprint arXiv:1603.00988*, 2016.

H. Montanelli and H. Yang. Error bounds for deep ReLU networks using the Kolmogorov–Arnold superposition theorem. *arXiv preprint arXiv:1906.11945*, 2019.

P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

P. Petersen, M. Raslan, and F. Voigtlaender. Topological properties of the set of functions generated by neural networks of fixed size. *arXiv preprint arXiv:1806.08459*, 2018.

Y. Qin. *Analytic Inequalities and Their Applications in PDEs*. Springer, 2017.

P. Ramachandran, B. Zoph, and Q. V. Le. Searching for Activation Functions. *arXiv preprint arXiv:1710.05941*, 2017.

C. Reisinger and Y. Zhang. Rectified deep neural networks overcome the curse of dimensionality for nonsmooth value functions in zero-sum games of nonlinear stiff systems. *arXiv preprint arXiv:1903.06652*, 2019.

F. Rousseau, L. Drumetz, and R. Fablet. Residual Networks as Flows of Diffeomorphisms. *Journal of Mathematical Imaging and Vision*, pages 1–11, 2019.

Z. Shen, H. Yang, and S. Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119: 74–84, 2019.

J. Takeuchi and Y. Kosugi. Neural network representation of finite element method. *Neural Networks*, 7(2):389–395, 1994.

M. Thorpe and Y. van Gennip. Deep Limits of Residual Neural Networks. *arXiv preprint arXiv:1810.11741*, 2018.

N. Yadav, A. Yadav, M. Kumar, et al. *An Introduction to Neural Network Methods for Differential Equations*. Springer, 2015.

D. Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. *arXiv preprint arXiv:1802.03620*, 2018.

L. Younes. *Shapes and Diffeomorphisms*, volume 171. Springer, 2010.

H. Zhang, X. Gao, J. Unterman, and T. Arodz. Approximation Capabilities of Neural Ordinary Differential Equations. *arXiv preprint arXiv:1907.12998*, 2019.

A Universal approximation with ReLU networks

This section is concerned with the proof of the universal approximation result in Proposition 4. Similar proofs relying on the approximation through interpolation can be found in Hanin (2017); Yarotsky (2018). We also follow He et al. (2018) for the expression of nodal basis functions, however, we also bound the complexity of the ReLU networks needed to express such functions.

A.1 Triangulations and piecewise linear functions

Let in the following $\mathcal{T}$ be a *locally finite triangulation* of the entire Euclidean space $\mathbb{R}^d$ consisting of nondegenerate $d+1$ simplices $\{\tau_k\}_{k \in \mathbb{N}}$ and vertices $\mathcal{V}$. More precisely, this means that the union of the simplices covers the entire space but that their interiors are pairwise disjoint and that every bounded set only intersects with finitely many simplices. Further, every simplex should be the convex hull of $d+1$ points and have non trivial interior.

For a vertex $x \in \mathcal{V}$ we set $N(x) \coloneqq \{k \in \mathbb{N} \mid x \in \tau_k\}$ define the *maximum number of neighboring simplices* to be

$$k_{\mathcal{T}} \coloneqq \sup_{x \in \mathcal{V}} |N(x)|$$

which we will assume to be finite. Further, we set

$$\Omega(x) \coloneqq \bigcup_{k \in N(x)} \tau_k$$

and call $\mathcal{T}$ *locally convex*, if $\Omega(x)$ is convex for all $x \in \mathcal{V}$. The *fineness* of the triangulation is defined to be the supremum over the diameters of the simplices

$$|\mathcal{T}| \coloneqq \sup_{k \in \mathbb{N}} \operatorname{diam}(\tau_k)$$

and we will assume that is finite. We will later give an explicit construction of a triangulation that satisfies those conditions.

Definition 5 (Piecewise linear functions)**.** We say a function $f \colon \mathbb{R}^d \to \mathbb{R}$ is *piecewise linear (PWL) with respect to* $\mathcal{T}$ if it is affine linear on every simplex of the triangulation. Given such a function f we call

$$|\mathcal{V}(f)| \coloneqq |\{x \in \mathcal{V} \mid f(x) \neq 0\}|$$

the *degrees of freedom* of the function.

Note that the definition of PWL functions automatically implies continuity since the affine regions are closed and cover $\mathbb{R}^d$ and affine functions are continuous. It is well known from the theory of finite elements that for every vertex $x \in \mathcal{V}$ there is a with respect to $\mathcal{T}$ piecewise linear function ϕ that satisfies $\phi(x) = 1$ and vanishies at every other vertex. We call this function the *nodal basis function* associated with x. The nodal basis functions form a basis of the space of PWL functions with finitely many degrees of freedom.

We will give an explicit construction of a triangulation that satisfies the assumptions from above. For this note that the unit cube $[0,1]^d$ can be divided into the simplices

$$S_\sigma \coloneqq \Big\{ x \in \mathbb{R}^d \;\Big|\; 0 \le x_{\sigma(1)} \le \dots \le x_{\sigma(d)} \le 1 \Big\}$$

where σ is a permutation of the set $\{1, \dots, d\}$. It is straight forward to check that those simplices cover the unit cube and have disjoint interiors and are non degenerate $d+1$ simplices. The fineness of this triangulation is $\sqrt{d}$. We call this triangulation the *standard triangulation* of the Euclidean space $\mathbb{R}^d$.

We will need the fact that the standard triangulation is locally convex. Since it is periodic, it suffices to show that $\Omega(0)$ is convex. In order to do this we will show that

$$\Omega(0) = \Big\{ z \in [-1,1]^d \;\Big|\; z_i \le z_j + 1 \text{ for all } i, j = 1, \dots, d \Big\} \eqqcolon A.$$

This expresses $\Omega(0)$ as an intersection of convex sets and hence shows the convexity of $\Omega(0)$.

Let us take $z = x - y \in \Omega(0)$ with $x, y \in S_\sigma$ where y has binary entries. Then we obviously have $z \in [-1,1]^d$. Let now $i, j \in \{1,\dots,d\}$, then we have to distinguish two cases. The first one is $\sigma^{-1}(i) \le \sigma^{-1}(j)$ which implies $x_i \le x_j$ and $y_i \le y_j$ and thus

$$z_i - z_j = (x_i - x_j) + (y_j - y_i) \le y_j \le 1.$$

For $\sigma^{-1}(i) > \sigma^{-1}(j)$ an analogue computation shows $z_i \le z_j + 1$ and hence we obtain the inclusion $\Omega(0) \subseteq A$. To see that the other inclusion holds true, we fix $z \in A$ and set $I \coloneqq \{i \mid z_i \ge 0\}$ and $J \coloneqq \{1,\dots,d\} \setminus I$. Further, we define $y \in \{0,1\}^d$ via

$$y_i \coloneqq \begin{cases} 0 & \text{for } i \in I \\ 1 & \text{otherwise} \end{cases}$$

and $x \coloneqq z + y$. By construction we have $z = x - y$ and $x, y \in [0,1]^d, y \in \{0,1\}^d$ and hence we only need to show the existence of a permutation σ such that $x, y \in S_\sigma$. Obviously, the statement $y \in S_\sigma$ is equivalent to $\sigma^{-1}(i) \le \sigma^{-1}(j)$ for all $i \in I, j \in J$. Since for $i \in I$ and $j \in J$ we have

$$x_i = z_i \le z_j + 1 = x_j,$$

there is a permutation that additionally satisfies $\sigma^{-1}(i) \le \sigma^{-1}(j)$ whenever $x_i \ge x_j$ for some $i, j \in \{1,\dots,d\}$.

A.2 Exact expression of piecewise linear functions as ReLU networks

We quickly present well known examples of functions that can exactly be expressed by ReLU networks (see He et al., 2018; Petersen and Voigtlaender, 2018).

1. *Identity mapping.* A basic calculation shows the identity

$$x = \rho(x) - \rho(-x) \quad \text{for all } x \in \mathbb{R}^d. \tag{4}$$

Hence, the identity is can be expressed as a ReLU network of width $2d$ which is visualised below. Note, that one can express the identity function as arbitrarily deep ReLU networks of width $2d$ since one can simply add more hidden layers where the affine linear transformation is the identity.

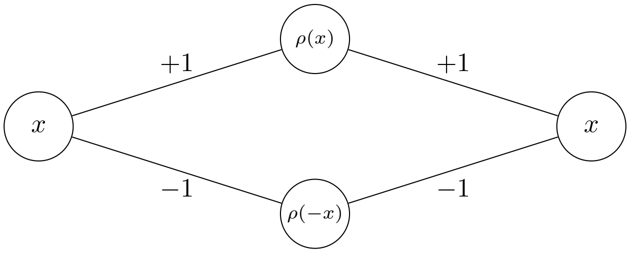

Figure 1: An example for an expression of the identity mapping as a ReLU network.

Similarly, one obtains that the absolute value can be expressed as a ReLU network of arbitrary depth and width 2 since $|x| = \rho(x) + \rho(-x)$.

2. *Minimum operation.* It is elementary to check

$$\min(x, y) = \frac{1}{2}\big(x + y - |x - y|\,\big).$$

We have already seen how the terms on the right hand side can be expressed as shallow ReLU networks and hence we obtain

$$\min(x, y) = \frac{1}{2}\Big(\rho(x + y) - \rho(-x - y) - \rho(x - y) - \rho(-x + y)\Big). \tag{5}$$

Therefore, the minimum operation can be expressed as a shallow ReLU network of width 4 and with weights $\pm\frac{1}{2}, \pm 1$.

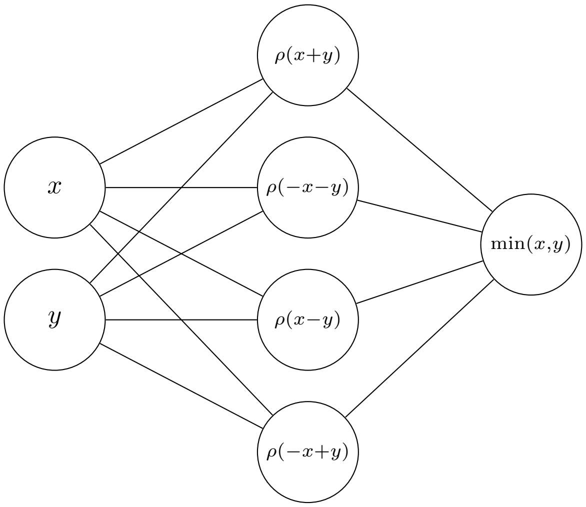

Figure 2: Expressing the minimum operation through a shallow ReLU network.

A consequence of the fact that the identity can be expressed as a shallow network is that the class of neural networks is closed under summation and also parallelisation. We will use those concepts of parallelisation and summation of networks which are relatively intuitive and we refer to Petersen and Voigtlaender (2018) for further details.

The expression of nodal basis functions as ReLU networks relies on the following proposition.

Lemma 6. *Let $\mathcal{T}$ be a locally finite and locally convex triangulation of $\mathbb{R}^d$ and let $x \in \mathcal{V}$ with nodal basis function ϕ. Then we have*

$$\phi(y) = \max\left\{0, \min_{k\in N(x)} g_k(y)\right\} = \min_{k\in N(x)} \rho(g_k(y)) \quad \textit{for all } y \in \mathbb{R}^d, \tag{6}$$

where g_k is the globally affine linear function that agrees with ϕ on the simplex τ_k.

For a proof we refer to He et al. (2018) which we also follow closely for the next two results, however, we additionally bound the complexity of the neural networks.

Proposition 7 (Minimum function)**.** *The minimum function $\min\colon \mathbb{R}^d \to \mathbb{R}$ can be expressed through a ReLU network of depth $\lceil \log_2(d) \rceil + 1$. Further, such a network can be constructed with weights $\left\{0, \pm\frac{1}{2}, \pm 1\right\}$ and $\mathcal{O}(d)$ many neurons and $\mathcal{O}(d)$ non-zero weights.*

Proof. Let us for the sake of easy notation assume $d = 2^m$. The construction of the representation of the minimum function relies on the observation that the minimum operation is the composition of $\log_2(d) = m$ mappings of the form

$$f_k\colon \mathbb{R}^{2^k} \to \mathbb{R}^{2^{k-1}}, \qquad \begin{pmatrix} x_1 \\ \vdots \\ x_{2^k} \end{pmatrix} \mapsto \begin{pmatrix} \min(x_1, x_2) \\ \vdots \\ \min(x_{2^k-1}, x_{2^k}) \end{pmatrix}.$$

Those functions are the realisation of a parallelisation of the representation of the minimum function constructed in Example 5. More precisely, f_k can be represented through a shallow ReLU network where the dimension of the hidden layer is $2 \cdot 2^k$. The concatenation of the m networks that represent the functions f_k is a representation of the minimum function of depth $m + 1$. By adding the dimensions of the layers we obtain the this network has

$$2^m + 2 \cdot 2^m + \cdots + 2^2 + 1 = 5d - 3$$

neurons and $4 \cdot (5d - 4)$ non-zero weights. □

Theorem 8 (Exact expression of PWL functions as ReLU networks)**.** *Consider $d, m \in \mathbb{N}$ and let $\mathcal{T}$ be a locally finite and locally convex triangulation of $\mathbb{R}^d$ with $k_{\mathcal{T}} < \infty$. Every function*

$f \colon \mathbb{R}^d \to \mathbb{R}^m$ that is piecewise linear with respect to $\mathcal{T}$ with N degrees of freedom can be expressed as a deep ReLU network with depth $\lceil \log_2(k_\mathcal{T}) \rceil + 2$ and at most $\mathcal{O}(mk_\mathcal{T}N + d)$ neurons. Further, all but $m(d+1)k_\mathcal{T}N$ weights can be fixed.

Proof. We assume $m = 1$ and note that the general statement follows from building a parallelised network. Since f is the linear combination of N nodal basis functions ϕ and hence it suffices to represent ϕ through a neural network as a representation of f can be obtained by considering the standard addition of those networks.

In order to represent ϕ, we use (6) and the previous proposition. For the sake of easy notation we assume that $N(x) = \{1, \dots, M\}$, then ϕ can be represented by the following network depicted in Figure 3 where the dashed part stands for a representation of the minimum function. It is clear that all weights except the ones of the first layer – which are $(d+1)M \leq (d+1)k_\mathcal{T}$ many – are fixed. □

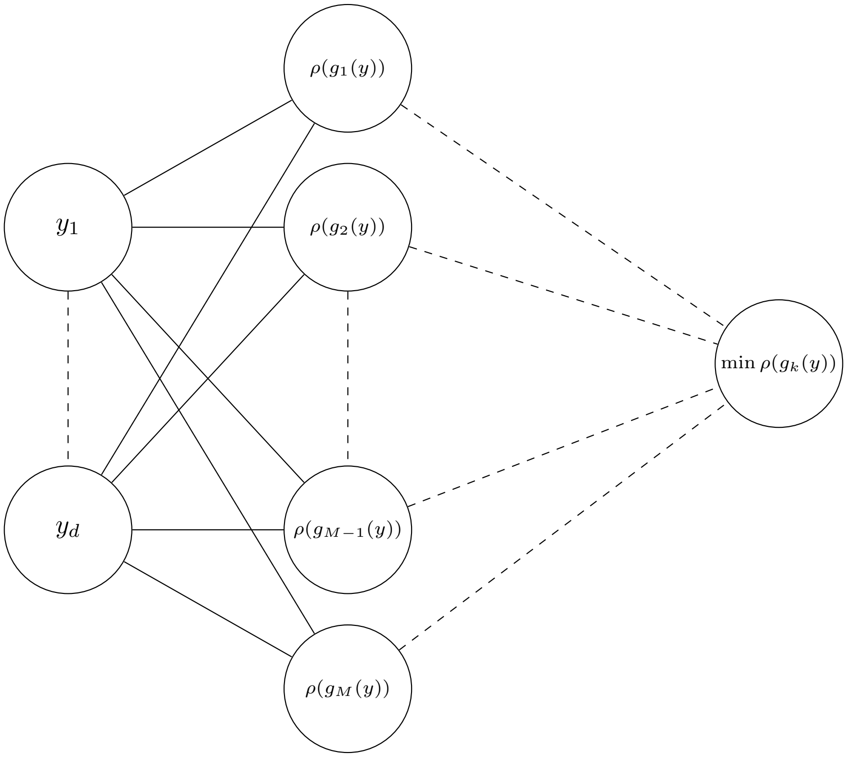

Figure 3: An expression of a nodal basis function ϕ where the dashed part stands for the expression of the minimum function that was constructed in Proposition 7.

A.3 Universal approximation through interpolation

We introduce the *modulus of continuity*

$$w_f \colon [0, \infty) \to [0, \infty], \quad \delta \mapsto \sup \Big\{ \|f(x) - f(y)\| \;\Big|\; x, y \in \Omega, \|x - y\| \leq \delta \Big\}.$$

It is elementary to check that a function is uniformly continuous if and only if the modulus of continuity takes finite values and is continuous. The pseudoinverse $w_f^{-1} \colon [0, \infty) \to [0, \infty)$ of the modulus of continuity is defined via

$$\varepsilon \mapsto \inf \{ \delta > 0 \mid w_f(\delta) > \varepsilon \}.$$

Note that if w_f is continuous, we have $w_f(w_f^{-1}(\varepsilon)) = \varepsilon$ for all $\varepsilon > 0$, i.e. we have

$$\|f(x) - f(y)\| \leq \varepsilon \quad \text{for all } x, y \in \Omega \text{ with } \|x - y\| \leq w_f^{-1}(\varepsilon).$$

Finally, if f is Lipschitz continuous with constant L we have $w_f(\delta) \leq L\delta$ and hence $w_f^{-1}(\varepsilon) \geq \frac{\varepsilon}{L}$.

Proposition 9 (Function approximation with piecewise linear functions)**.** *Let $d, m \in \mathbb{N}$ and $f\colon \mathbb{R}^d \to \mathbb{R}^m$ be a continuous function and $\mathcal{T}$ be a locally finite triangulation of the Euclidean space $\mathbb{R}^d$ with fineness $\delta \in (0, \infty)$. Let $\Omega \subseteq \mathbb{R}^d$ be a union of simplices of $\mathcal{T}$ and g be the with respect to $\mathcal{T}$ piecewise linear function that agrees with f on all vertices inside of Ω and vanishes everywhere else. Then we have*

$$\|f - g\|_{\infty,\Omega} \leq w_{f|\Omega}(\delta).$$

Finally, we have $\|g\|_\infty \leq \|f\|_\infty$.

Proof. Let $x \in \Omega$ then x lies in a convex simplex with vertices $x_1, \dots, x_{d+1} \in \Omega$. Hence, we find convex weights $\alpha_1, \dots, \alpha_{d+1} \in [0, 1]$ such that $x = \sum_{i=1}^{d+1} \alpha_i x_i$. Now we obtain

$$\big\|f(x) - g(x)\big\| = \left\|f(x) - \sum_{i=1}^{d+1} \alpha_i f(x_i)\right\| \leq \sum_{i=1}^{d+1} \alpha_i \big\|f(x) - f(x_i)\big\| \leq w_{f|\Omega}(\delta).$$

Furthermore, if $\|f\|$ is bounded by c, then we obtain

$$\big\|g(x)\big\| \leq \sum_{i=1}^{d+1} \alpha_i \cdot \big\|f(x_i)\big\| \leq \sum_{i=1}^{d+1} \alpha_i \cdot c = c$$

for all $x \in \mathbb{R}^d$. □

Combining the previous results with the construction of the standard triangulation we obtain the following result.

Proposition 10 (Universal approximation with ReLU networks)**.** *Consider a continuous function $f\colon \mathbb{R}^d \to \mathbb{R}^m$ where $d, m \in \mathbb{N}$. Let further $r > 0$ and $\varepsilon > 0$ and let $w_{f,r}$ be the modulus of continuity of $f|_{[-r,r]^d}$. Then for every $\varepsilon > 0$ there is a ReLU network R_ε with parameters θ_ε that satisfies the following:*

1. Approximation: *It holds that* $\sup_{x\in[-r,r]^d} \|f(x) - R_\varepsilon(x)\| \leq \varepsilon$.
2. Complexity bounds: *The network has depth* $\lceil \log_2((d + 1)!)\rceil + 2$ *,* $\mathcal{O}\big(\omega_{f,r}^{-1}(\varepsilon)^{-d}\big)$ *many neurons and all but* $\mathcal{O}\big(\omega_{f,r}^{-1}(\varepsilon)^{-d}\big)$ *weights can be fixed. Finally, we have* $\|R_\varepsilon\|_\infty \leq \|f\|_\infty$.

Proof. Building on the previous results we only have to check that there is a triangulation $\mathcal{T}$ with fineness at most $w_{f,r}^{-1}(\varepsilon)$, $k_\mathcal{T} < \infty$[4] such that $[-r, r]^d$ is the union of simplices and

$$2^d \cdot \left\lceil \frac{\sqrt{d} \cdot r}{w_{f,r}^{-1}(\varepsilon)} \right\rceil^d$$

vertices in $[-r, r]^d$. We obtain this triangulation by scaling the standard triangulation by

$$r \cdot \left\lceil \frac{\sqrt{d} \cdot r}{w_{f,r}^{-1}(\varepsilon)} \right\rceil^{-1},$$

for which the properties are easily verified. □

This result can easily be rewritten for Lipschitz continuous functions as the Lipschitz continuity controls the modulus of continuity. We obtain the following approximation result.

Proposition 4 (Universal approximation under Lipschitz condition)**.** *Let $d, m \in \mathbb{N}$ and $r > 0$ and let $f\colon \mathbb{R}^d \to \mathbb{R}^m$ be Lipschitz continuous. Then for every $\varepsilon > 0$ there is a ReLU network R_ε with parameters θ_ε that satisfies the following:*

1. Approximation: *It holds that* $\sup_{x\in[-r,r]^d} \|f(x) - R_\varepsilon(x)\| \leq \varepsilon$.
2. Complexity bounds: *The network has depth* $\lceil \log_2((d{+}1)!)\rceil{+}2$ *,* $\mathcal{O}\left(r^d\varepsilon^{-d}\right)$ *many neurons and all but* $\mathcal{O}\left(r^d\varepsilon^{-d}\right)$ *weights can be fixed. Finally, if $\|f\|$ is bounded by c so is $\|R_\varepsilon\|$.*

[4] One can count the neighboring points and show $k_\mathcal{T} = (d + 1)!$.

B Error estimate for perturbed Euler schemes

First, we need to introduce the notion of weak solutions of ordinary differential equations.

Definition 11 (Weak solutions)**.** Let $f\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ be a Carathéodory function, i.e., measurable in the first and continuous in the second argument and let further $x_0 \in \mathbb{R}^d$. Then we say $x\colon [0,1] \to \mathbb{R}^d$ is a *weak solution* of the differential equation

$$\partial_t x(t) = f(t, x(t)), \quad x(0) = x_0$$

if it satisfies

$$x(t) = x_0 + \int_0^t f(s, x(s))\mathrm{d}s \quad \text{for all } t \in [0,1].$$

The integral on the right hand side can be interpreted as a componentwise Lebesgue integral where the Carathéodory condition ensures the measurability. Further, we call $x\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ the *space-time solution* of the ODE with right hand side f if it solves

$$\partial_t x(t,y) = f(t, x(t,y)), \quad x(0,y) = y.$$

The well posedness of ordinary differential equations in the weak sense can be proved just like the well posedness results from the classical theory. In particular, a global solution $x\colon [0,1] \to \mathbb{R}^d$ exists for every initial value $x_0 \in \mathbb{R}^d$ if $f(t,\cdot)$ is bounded and Lipschitz continuous for almost all t with integrable uniform norm and Lipschitz constant (see Younes, 2010). We denote the space of those functions which are also Bochner-measurable[5] by $L^1([0,1]; \mathcal{C}_b^{0,1}(\mathbb{R}^d;\mathbb{R}^d))$.

Definition 12 (Euler discretisation)**.** Let $0 = t_0 < \cdots < t_n = 1$ be a partition of the unit interval and $x_0 \in \mathbb{R}^d$. Let $f\colon [0,1]\times\mathbb{R} \to \mathbb{R}$ be an arbitrary Carathéodory function. Then we define the *Euler discretisation or Euler scheme to the right hand side f, initial value x_0 and with respect to the partition* $(t_0, t_1, \ldots, t_n)$ via

$$x^n(0) \coloneqq x_0, \quad \text{and } x^n(t_{i+1}) = x^n(t_i) + (t_{i+1} - t_i) f(t_i, x^n(t_i))$$

and linearly in between.

It is important to note that the Euler discretisation x^n satisfies the integral equation

$$x^n(t) = x_0 + \int\limits_0^t \gamma(s)\mathrm{d}s \quad \text{for all } t \in [0,1],$$

where

$$\gamma(t) \coloneqq \sum_{i=0}^{n-1} \chi_{[t_i, t_{i+1})} f(t_i, x(t_i)).$$

Lemma 13 (Generalised Grönwall's inequality)**.** *Let $x_0, y_0 \in \mathbb{R}^d$ and let $\gamma_0, \gamma_1 \in L^1([0,1];\mathbb{R}^d)$[6] and let x and y satisfy the integral equations*

$$x(t) = x_0 + \int\limits_0^t \gamma_1(s)\mathrm{d}s \quad \textit{and } y(t) = y_0 + \int\limits_0^t \gamma_2(s)\mathrm{d}s \quad \textit{for all } t \in [0,1].$$

Assume now that there are non negative functions $\alpha, \beta \in L^1([0,1]])$ such that

$$\|\gamma_1(t) - \gamma_2(t)\| \le \alpha(t) + \beta(t)\cdot\|x(t) - y(t)\| \quad \textit{for all } t \in [0,1].$$

Then we have

$$\|x(t) - y(t)\| \le c\cdot\Big(\|x_0 - y_0\| + \|\alpha\|_{L^1(I)}\Big) \quad \textit{for all } t \in [0,1],$$

where we can choose

$$c = 1 + \|\beta\|_{L^1([0,1])}\cdot\exp(\|\beta\|_{L^1([0,1])}).$$

[5] See Diestel and Uhl (1977); there such functions are called strongly measurable.

[6] i.e., their norms are integrable; see Diestel and Uhl (1977) for an introduction to vector valued integration.

Proof. For $t \geq t_0$ we compute

$$\begin{aligned}\|x(t)-y(t)\| \leq&\ \|x_0-y_0\| + \int\limits_{t_0}^{t} \|\gamma_1(s)-\gamma_2(s)\|\,\mathrm{d}s \\ \leq&\ \|x_0-y_0\| + \int\limits_{t_0}^{t} \alpha(s)\mathrm{d}s + \int\limits_{t_0}^{t} \beta(s)\cdot\|x(s)-y(s)\|\,\mathrm{d}s.\end{aligned}$$

An application of Grönwall's inequality yields the assertion.[7] For $t \leq t_0$ the computation follows in analogue way or by reflection. □

Remark 14. If $\|f\|$ is bounded by c we obtain the growth estimate

$$\|x(t)\| \leq \|x_0\| + c.$$

Further, this estimate holds also for all Euler discretisations of f.

Proposition 15 (Continuity of solution map)**.** *Let $x_0, y_0 \in \mathbb{R}^d$ and let $f, g\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ be Carathéodory functions such that $f(t,\cdot)$ is Lipschitz continuous with constant $h(t)$ for $t \in [0,1]$ where $h \in L^1([0,1])$. Further, let $f-g \in L^1([0,1]; L^\infty(\mathbb{R}^d;\mathbb{R}^d))$[8] and let $x, y\colon [a,b] \to \mathbb{R}^d$ be weak solutions to the differential equations*

$$\partial_t x(t) = f(t, x(t)), \quad x(t_0) = x_0 \quad \textit{and}\ \partial_t y(t) = g(t, y(t)), \quad y(t_0) = y_0.$$

Then we have

$$\sup_{t\in[0,1]} \|x(t)-y(t)\| \leq c \cdot \left(\|x_0-y_0\| + \|f-g\|_{L^1([0,1];L^\infty(\mathbb{R}^d;\mathbb{R}^d))}\right), \tag{7}$$

where the constant c only depends on $\|h\|_{L^1([0,1])}$.

Proof. We only need to check the requirements of the previous result. We recall that x and y solve the integral equations associated to the ODEs and hence obtain for $t \in I$

$$\begin{aligned}\|\gamma_1(t)-\gamma_2(t)\| =&\ \|f(t,x(t)) - g(t,y(t))\| \\ \leq&\ \|f(t,x(t)) - f(t,y(t))\| + \|f(t,y(t)) - g(t,y(t))\| \\ \leq&\ h(t)\cdot\|x(t)-y(t)\| + \|f(t,\cdot)-g(t,\cdot)\|_\infty .\end{aligned}$$

□

Later we will perceive residual networks as an perturbed Euler approximation of an ordinary differential equation. To show convergence of those we provide an error estimate for such perturbations, namely we replace the direction $f(t_i, x^n(t_i))$ of the Euler approximation x^n on $[t_i, t_{i+1})$ by $z_i \approx f(t_i, x^n(t_i))$.

Proposition 16 (Error estimate for perturbed Euler schemes)**.** *Let $f\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ be a Carathéodory function such that $f(t,\cdot)$ is Lipschitz with constant $h(t)$ where $h \in L^1([0,1])$. Let now $x_0 \in \mathbb{R}^d$ and $x\colon [0,1] \to \mathbb{R}^d$ be the weak solution to*

$$\partial_t x(t) = f(t, x(t)) \quad \textit{and}\ x(0) = x_0.$$

Fix $z_0, \ldots, z_{n-1} \in \mathbb{R}^d$ and set $t_i := i/n$ as well as $\gamma := \sum_{i=0}^{n-1} \chi_{[t_i,t_{i+1})} z_i$. Let $x^n\colon [0,1] \to \mathbb{R}^d$ satisfy the integral equation

$$x^n(t) = x_0 + \int\limits_0^t \gamma(s)\mathrm{d}s \quad \textit{for all}\ t \in [0,1].$$

Assume that we have $\|z_i - f(t, x^n(t_i))\| \leq \varepsilon$ for all $t \in [t_i, t_{i+1}), i = 0, \ldots, n-1$ as well as $\|z_i\| \leq c$ for all $i = 0, \ldots, n-1$. Then we have

$$\|x^n(t) - x(t)\| \leq \tilde{c}\cdot\left(\varepsilon + \frac{c}{n}\cdot\|h\|_{L^1([0,1])}\right),$$

where $\tilde{c}$ only depends on $\|h\|_{L^1([0,1])}$.

[7] For a general version of Grönwall's inequality we refere to Theorem 1.2.8 in Qin (2017).

[8] i.e., the uniform distance $\|f(t,\cdot)-g(t,\cdot)\|_\infty$ is integrable over $[0,1]$.

Proof. Once more we will use Lemma 13 with obvious choices of γ_1 and γ_2. For $t \in [t_i, t_{i+1})$ we estimate

$$\begin{aligned}\|\gamma_1(t) - \gamma_2(t)\| &= \|z_i - f(t, x(t))\| \\ &\leq \|z_i - f(t, x^n(t_i))\| + \|f(t, x^n(t_i)) - f(t, x(t))\| \\ &\leq \varepsilon + \|f(t, x^n(t_i)) - f(t, x^n(t))\| + \|f(t, x^n(t)) - f(t, x(t))\| \\ &\leq \varepsilon + h(t) \cdot \|x^n(t_i) - x^n(t)\| + h(t) \cdot \|x^n(t) - x(t)\| \\ &\leq \varepsilon + \frac{c}{n} \cdot h(t) + h(t) \cdot \|x^n(t) - x(t)\| .\end{aligned}$$

□

C Proofs of the main results

Let us quickly recall our definition of residual networks. Let $R_1, \ldots, R_n \colon \mathbb{R}^d \to \mathbb{R}^d$ be neural networks. The resulting ResNet $x^n \colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ is defined via

$$x^n(0, y) \coloneqq y, \quad x^n(t_{k+1}, y) \coloneqq x^n(t_k, y) + n^{-1} \cdot R_{k+1}(x^n(t_k, y))$$

for $k = 0, \ldots, n-1$ and linearly in between.

It is clear from the definition that ResNets are in fact Euler approximations to the piecewise constant right hand side $f \colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ which is defined via

$$f(t, x) \coloneqq \sum_{i=0}^{n-1} \chi_{[t_i, t_{i+1})}(t) R_{i+1}(x) \quad \text{for all } t \in [0,1], x \in \mathbb{R}^d,$$

where $t_i \coloneqq i/n$. We will use the error estimate on perturbed Euler schemes established in the previous chapter to show that by letting $n \to \infty$ and increasing the expressivity of the networks R_i ReLU ResNets are able to approximate space-time solutions of arbitrary right hand sides.

Lemma 17. *Let $f \in L^1([0,1]; \mathcal{C}_b^{0,1}(\mathbb{R}^d; \mathbb{R}^d))$ and let x be the space-time solution of the ODE with right hand side f. Then for every $\varepsilon > 0$ there is $n \in \mathbb{N}$ and $g \in L^1([0,1]; \mathcal{C}_b^{0,1}(\mathbb{R}^d; \mathbb{R}^d))$ that is constant on all intervals of the form $[i/n, (i+1)/n)$ such that the space-time solution $\tilde{x}$ to g satisfies*

$$\|x(t, y) - \tilde{x}(t, y)\| \leq \varepsilon \quad \textit{for all } t \in [0,1], y \in \mathbb{R}^d.$$

Proof. By standard Bochner theory (see Arendt et al., 2011) the continuous functions are dense in

$$L^1\left([0,1]; \mathcal{C}_b^{0,1}\left(\mathbb{R}^d; \mathbb{R}^d\right)\right).$$

However, continuous functions can approximated arbitrarily well by functions that are constant on intervals of equal length. Now the continuity estimate (7) yields the assertion. □

Theorem 2 (Space-time approximation with ResNets)**.** *Let $d \in \mathbb{N}$ and*

$$f \in L^1\left([0,1]; \mathcal{C}_b^{0,1}\left(\mathbb{R}^d; \mathbb{R}^d\right)\right)$$

and let x be the space-time solution to f. Then for every compact set $K \subseteq \mathbb{R}^d$ and $\varepsilon > 0$ there is a ReLU ResNet $\tilde{x}$ such that

$$\|\tilde{x}(t, y) - x(t, y)\| \leq \varepsilon \quad \textit{for all } t \in [0,1], y \in K.$$

Proof. By the previous lemma we can without loss of generality assume that f is constant on the intervals $[i/n, (i+1)/n)$ for some $n \in \mathbb{N}$. We note that since f is piecewise constant with values in $\mathcal{C}_b^{0,1}(\mathbb{R}^d; \mathbb{R}^d)$ there is $c > 0$ such that

$$\|f(t, x)\| \leq c \quad \text{for all } t \in [0,1], x \in \mathbb{R}^d.$$

It suffices to show the statement for the compact set $K = \overline{B_N}$ where B_N denotes the ball of radius N around the origin. By Remark 14 we have $x(t, y) \in \overline{B_M}$ for every $t \in [0,1], y \in \overline{B_N}$ where $M = N + c$.

Let now $\varepsilon > 0$, then the universal approximation result 10 for ReLU networks yields the existence of ReLU networks $R_0, \dots, R_{n-1} \colon \mathbb{R}^d \to \mathbb{R}^d$ with parameters $\theta_0, \dots, \theta_{n-1}$ such that

$$\|f(t,y) - R_i(y)\| \leq \varepsilon \quad \text{for all } y \in \overline{B_M}, t \in [i/n, (i+1)/n). \tag{8}$$

as well as $\|R_i\| \leq c$. Further, we choose $k \in \mathbb{N}$ such that

$$\frac{c}{kn} \cdot \|f\|_{L^1([0,1];\mathcal{C}_b^{0,1}(\mathbb{R}^d;\mathbb{R}^d))} \leq \varepsilon. \tag{9}$$

Let now $\tilde{x}$ be the ReLU ResNet with parameters

$$(\theta_0, \dots, \theta_0, \theta_1, \dots, \theta_1, \dots, \theta_{n-1}, \dots, \theta_{n-1}), \tag{10}$$

where each network θ_i is included k times. Now we aim to apply Proposition 16 and hence check its requirements and fix $y \in \overline{B_N}$ and denote $x(t,y), \tilde{x}(t,y)$ with $x(t)$ and $\tilde{x}(t)$ respectively and again Remark 14 yields $\tilde{x}(t) \in \overline{B_M}$ for all $t \in [0,1]$. In order to use the notation from the proposition we set $t_i \coloneqq i/(kn)$ and $z_i \coloneqq R_j(\tilde{x}(t_i))$ for $i = kj, \dots, k(j+1) - 1$ and obtain

$$\tilde{x}(t) = y + \int_0^t \gamma(s) \mathrm{d}s \quad \text{for } \gamma = \sum_{i=0}^{kn-1} \chi_{[t_i, t_{i+1})} z_i.$$

Further, it holds that

$$\|z_i - f(t, \tilde{x}(t_i))\| \leq \varepsilon \quad \text{for all } t \in [t_i, t_{i+1}), i = 0, \dots, kn-1$$

as well as $\|z_i\| \leq c$. Now Proposition 16 yields

$$\begin{aligned} \|\tilde{x}(t) - x(t)\| &\leq \tilde{c} \cdot \left(\varepsilon + \frac{c}{kn} \cdot \|f\|_{L^1([0,1];\mathcal{C}_b^{0,1}(\mathbb{R}^d;\mathbb{R}^d))}\right) \\ &\leq 2\tilde{c} \cdot \varepsilon \quad \text{for all } t \in [0,1], \end{aligned}$$

where $\tilde{c}$ only depends on $\|f\|_{L^1([0,1];\mathcal{C}_b^{0,1}(\mathbb{R}^d;\mathbb{R}^d))}$ and not on $y \in \overline{B_N}$. □

The universal approximation theorem presented above is of qualitative nature since it does not give any estimates on the complexity of the residual network needed to approximate a flow up to a certain precision. This is due to the fact that we work with density results for continuous functions in the Bochner space $L^1([0,1]; \mathcal{C}_b^{0,1}(\mathbb{R}^d; \mathbb{R}^d))$. In the proof above one could also assume that (9) holds for $k = 1$ since f is also piecewise constant on the intervals $[i/(kn), (i+1)/(kn))$. However, we wanted to separate the approximation procedures in space and in time. More precisely, if f is (almost) constant in time, (8) can be achieved with little n and hence the constructed ResNet (2) shares a lot of weights. This observation could be used to explore approximation capabilities of ResNets with shared weights under different spatial and temporal regularity of the right hand side f.

We use analogue arguments to establish estimates on the number and complexity of residual blocks required to approximate space-time solutions of ODEs with Lipschitz continuous right hand side f.

Theorem 3 (Space-time approximation with complexity bounds)**.** *Let $d \in \mathbb{N}$, $(r_n)_{n\in\mathbb{N}} \subseteq (0,\infty)$ be a sequence convergent to ∞ and let $f \colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ be a bounded and Lipschitz continuous function. Let $x \colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ be the space-time solution of the ODE with right hand side f. Then for every $n \in \mathbb{N}$ there is a ReLU ResNet x^n with parameters $\theta^n = (\theta_1^n, \dots, \theta_n^n)$ such that the following are satisfied:*

1. Approximation: *For every compact set $K \subseteq \mathbb{R}$ it holds*

$$\sup_{t\in[0,1], y\in K} \|x^n(t,y) - x(t,y)\| \in \mathcal{O}(n^{-1}).$$

2. Complexity bounds: *Every residual block θ_k^n has depth $\lceil \log_2((d+1)!) \rceil + 2$ and satisfies*

$$N(\theta_k^n) \in \mathcal{O}\left(r_n^d n^d\right).$$

Finally, all but $\mathcal{O}\left(r_n^d n^d\right)$ weights can be fixed.

Proof. We fix $n \in \mathbb{N}$ and set $t_i \coloneqq i/n$. Let R_i^n be ReLU networks of asserted complexity that approximate $f(t_i, \cdot)$ on $[-r_n, r_n]^d$ up to n^{-1} which exist by Proposition 4. Let x^n be the ReLU ResNet with residual blocks $R_1^n, \dots, R_n^n$. We fix $N > 0$ and will show

$$\sup_{t \in [0,1], y \in \overline{B_N}} \|x^n(t, y) - x(t, y)\| \in \mathcal{O}(n^{-1}) \quad \text{for } n \to \infty$$

through an application of Proposition 16. Since f is bounded, there is $c > 0$ such that

$$\|f(t, y)\| \leq c \quad \text{for all } t \in [0, 1], y \in \mathbb{R}^d$$

and hence the functions R_i^n satisfy this as well. Setting $M \coloneqq N + c$, Remark 14 yields

$$x(t, y), x^n(t, y) \in \overline{B_M} \quad \text{for all } t \in [0, 1], y \in \overline{B_N}.$$

Now we fix $y \in \overline{B_N}$ and write $x(t), x^n(t)$ for $x(t, y)$ and $x^n(t, y)$ respectively; to keep to the notation of the error estimate for Euler schemes, we set $z_i \coloneqq R_i^n(x^n(t_i))$. For $n \geq M$ we obtain

$$\begin{aligned} \|z_i - f(t, x^n(t_i))\| &= \|R_i^n(x^n(t_i)) - f(t, x^n(t_i))\| \\ &\leq \|R_i^n(x^n(t_i)) - f(t_i, x^n(t_i))\| + \|f(t_i, x^n(t_i)) - f(t, x^n(t_i))\| \\ &\leq n^{-1} \cdot (1 + L) \end{aligned}$$

for all $t \in [t_i, t_{i+1})$ and $i = 0, \dots, n-1$ where L denotes the Lipschitz constant of f. Furthermore, we have $\|z_i\| \leq c$ for all $i = 0, \dots, n-1$ and hence Proposition 16 completes the proof. □