Adv Comput Math (2021) 47: 11
https://doi.org/10.1007/s10444-020-09834-7

Efficient approximation of solutions of parametric linear transport equations by ReLU DNNs

Fabian Laakmann[1] · Philipp Petersen[2]

Received: 5 February 2020 / Accepted: 15 December 2020 /
Published online: 28 January 2021

Abstract

We demonstrate that deep neural networks with the ReLU activation function can efficiently approximate the solutions of various types of parametric linear transport equations. For non-smooth initial conditions, the solutions of these PDEs are high-dimensional and non-smooth. Therefore, approximation of these functions suffers from a curse of dimension. We demonstrate that through their inherent compositionality deep neural networks can resolve the characteristic flow underlying the transport equations and thereby allow approximation rates independent of the parameter dimension.

Keywords Deep neural networks · Parametric PDEs · Approximation rates · Curse of dimension · Transport equations

Mathematics Subject Classification (2010) 35A35 · 35Q49 · 41A25 · 41A46 · 68T05 · 65N30

1 Introduction

Linear parametric transport equations play an essential role in engineering, modelling, and mathematical physics where they describe physical phenomena of heat and mass transfer. A typical example is the transport of pollution in air or water

Communicated by: Jan Hesthaven

✉ Philipp Petersen
Philipp.Petersen@univie.ac.at

Fabian Laakmann
Fabian.Laakmann@maths.ox.ac.uk

[1] Mathematical Institute, University of Oxford, Andrew Wiles Building, Woodstock Road, Oxford, OX2 6GG, UK

[2] Institut für Mathematik, Universität Wien, Kolingasse 14-16, 1090 Wien, Austria

depending on a set of parameters such as the direction and intensity of the flow of the fluid.

In this work, we study to what extent the solutions of various types of parametric linear transport equations can be efficiently represented by deep neural networks. Concretely, we study variations of the following problem: Let $n, D, k \in \mathbb{N}$, and $T > 0$. Let $V \in C^k([0, T] \times \mathbb{R}^n \times [0, 1]^D; \mathbb{R}^n)$, $f \in C^k([0, T] \times \mathbb{R}^n \times [0, 1]^D)$ and let $u_0 \in C^1(\mathbb{R}^n)$. We want to find $u \in C^1([0, T] \times \mathbb{R}^n \times [0, 1]^D)$ such that

$$\begin{cases} \partial_t u(t, x, \eta) + V(t, x, \eta) \cdot \nabla_x u(t, x, \eta) = f(t, x, \eta), \\ u(0, x, \eta) = u_0(x). \end{cases} \tag{1.1}$$

The PDE of (1.1) has been studied extensively (see, e.g., [2, 3, 10, 24]), and we will recall the fundamentals in Section 3. The setup that we have in mind is one where the dimension of the parameter space D is very high, V is smooth, and u_0 is not very regular. Hence, direct approximation of u amounts to *approximating a high-dimensional function of low regularity*. In this formulation, the task is extremely challenging for classical methods.

While the global approximation problem is almost intractable, the method of characteristics shows that, even though u is not smooth, its singularities revolve along smooth curves, called *characteristic curves*. In this framework, the function u can typically be written in a compositional form of two functions where one is high-dimensional and smooth and the other is low-dimensional and (potentially) rough.

Based on this split and the inherent compositionality of neural networks, we will demonstrate that every u satisfying (1.1) can be approximated efficiently by neural networks with ReLU activation function. The approximation rate is significantly better than that of any classical regularity-based approximation of u. In particular, in the prescribed setup, we will observe an *approximation rate independent of the dimension D of the parameter space*.

The material presented below was first established in a mini-project of the first author at the University of Oxford, [37].

1.1 Applications and relevance

We believe that the efficient approximation of solutions of parametric transport equations with dimension-independent approximation rates is an interesting and relevant problem in the following domains:

- *Approximation theory:* The class of functions that are solutions of high-dimensional parametric linear transport equations is a relevant but non-standard function class. This class while high-dimensional has a non-trivial but rigid structure imposed upon via the underlying PDE. It is, therefore, interesting to establish to what extent deep neural networks can leverage on this structure. In this context, similar approximation schemes based on structured systems were developed for special types of (parametric) transport equations. For example, in [15, 16] and [44] systems were introduced that approximate the solutions of linear transport equations with linear or C^2-regular characteristic curves. These constructions are closely tied to the type of characteristic curves. We demonstrate here that, in contrast to these systems, we do not choose specific types of deep neural networks depending on the specific problem to achieve the presented rates. Indeed, the only variable parameters are the depth and the width of neural networks. In that sense, deep neural networks can automatically adapt to the underlying regularity of the problems.

- *Estimation:* In machine learning and especially in deep learning, deep neural networks are trained with gradient-based optimization algorithms to minimize empirical energies based on random samples [26, 38]. These techniques have proven to be extremely successful in a variety of applications.

  Consider a parametric transport problem of the form (1.1) where u_0, f, and V are unknown, but samples $(u(t_i, x_i, \eta_i))_{i=1}^N$ are available through measurements. Such a scenario could be encountered in the transport of pollution in fluids under unknown circumstances, but with a control on parameters of the experiment. In this formulation, the transport problem is a standard supervised learning problem. Moreover, classical methods to solve linear transport equations cannot be used at all without knowledge of f and V in (1.1).

  Certainly, this estimation problem can only be successfully solved with deep learning techniques, if the correct solution to the problem can be represented or closely approximated by a deep neural network. In this context, our results show the feasibility of this approach.
- *Numerical analysis:* Deep neural networks have been employed as Ansatz spaces for PDEs in multiple settings before [5, 19, 55]. Of course, the efficiency of these methods depends on the capacity of the Ansatz space to capture the true solution.

  Our proposed approximation of solutions of (1.1) by deep neural networks can be thought of as a higher-order method that automatically adapts to the regularity of the underlying characteristic curves.

  An established method to solve (1.1) is by using (Petrov-) Galerkin-type discretizations [14, 20]. These methods are, however, typically not adaptive to singularities lying on lower-dimensional manifolds. In the model of (1.1), such structured singularities evolve precisely along the characteristic curves. While some advances to handle structured singularities have been made, e.g., [15], the adaptivity to the manifold only uses low-order information on the smoothness of the manifold. Our results demonstrate that an approximation via deep neural networks adapts to any regularity of the characteristic curves in the sense that the approximation quality improves for smoother characteristic curves. On the contrary, standard discretization and time-stepping techniques such as finite differences and Euler, Crank-Nicolson, or higher-order variants converge with rates depending on the global regularity of the solution of the PDE which, in our setup, is assumed to be quite low.
- *Reduced-order models:* In applications where the solution of (1.1) is requested for many different parameter values, it is desirable to employ model reduction techniques [31, 50]. It is well known that parametric linear transport problems are highly challenging for linear reduced-order models because the dimension of linear approximation spaces to capture the non-linear evolution of singularities can be excessive [45, Section 5] [17, Section 6.3]. Indeed, linear reduced-order

models typically succeed only if additional assumptions are made on the parameter dependence, such as a certain separability of the parametric dependence and the spatial dependence of V [27].

The neural network-based approximation presented in this work requires almost no structural assumption on the parametric dependence. Indeed, a smooth dependence of V and f on the parameters is sufficient. In this context, a superiority of deep neural network-based approaches over linear reduced-order models to solve parametric transport equations was also empirically observed in [23].

1.2 Related work

This work describes the capacity of neural networks to approximate high-dimensional functions with asymptotic rates independent of the underlying dimension. Of course, approximation theory of deep neural networks is a well-established field. Therefore, in order to place our contribution in the context of existing literature, we provide an overview of classical and more modern developments in the field below.

1.2.1 Classical approximation

The first and probably most prominent result describing the approximation capabilities of neural networks is the universal approximation theorem [13, 33]. This theorem states that, on a compact domain, every continuous function can be arbitrarily well approximated by a neural network in the uniform norm. These statements, however, do not provide an estimate on the required sizes of the approximating neural networks. The typical approach to obtain a quantitative estimate on the order of approximation is to re-approximate classical methods. For example, in [39] and [40], it was shown that neural networks yield the same approximation rates as splines when approximating smooth functions.

Recently, approximation by neural networks with the ReLU activation function has received the most attention since this activation function is arguably the most widely used in applications. It was demonstrated that deep neural networks with the ReLU activation function achieve the same approximation rates as linear and higher-order finite elements [30, 46], wavelets [54], and local approximation by Taylor polynomials [57].

These classical approximation results show that deep neural networks are very versatile by combining the approximation capabilities of a wide variety of classical tools. However, they do not identify a particular situation where deep neural networks outperform the best classical method. This picture changes drastically, when one considers high-dimensional approximation.

1.2.2 High-dimensional approximation

High-dimensional approximation generally suffers from a curse of dimension, meaning that approximation rates deteriorate exponentially with increasing dimension [7, 43]. Nonetheless, if an additional structure is assumed, then the curse of dimension can be overcome. It turns out that deep neural networks can take advantage of a wide variety of complex additional structural properties. For example, it was shown in [4] that deep neural networks can approximate functions with bounded first Fourier moments without a curse of dimension. Other regularity-based assumptions were used in [41] and [56]. Further classes of functions with structural assumptions such as functions based on directed acyclic graphs [49] or functions admitting strong invariances [48, Section 5] allow similar results. Finally, if the approximation error is evaluated on a low-dimensional manifold only, then [9, 11, 51, 54] show approximation rates independent of the ambient dimension.

1.2.3 Approximation of solutions of PDEs

The extraordinary efficiency in the approximation of certain high-dimensional functions has been especially interesting in connection with the numerical solution of PDEs [5, 19, 55]. For example, for high-dimensional Black Scholes-, Kolmogorov-, or heat equations, deep neural networks can efficiently approximate the solutions thereof in a regime where any mesh-based method would fail [6, 8, 21, 34]. In these works, a compositional structure of the solution of a PDE is derived via the Feynman-Kac formula. The approach via the method of characteristics of our work can be interpreted as a special case of the approach via the Feynman-Kac formula.

Moreover, in the framework of parametric problems, high-dimensional problems can be efficiently represented if there exist suitable representations thereof in a general reduced basis [36], or as a polynomial chaos expansion [47, 52].

1.3 Outline

In Section 2, we introduce all notions and fundamental results associated with neural networks. Section 3 is devoted to the introduction of various types of linear transport equations. In Section 4, we present the four main results of this work: Theorems 4.3, 4.4, 4.6, and 4.8. These results describe approximation rate bounds for the solutions of the equations of Section 3 by deep neural networks. Finally, in Section 5, we discuss natural extensions of the presented results. Some auxiliary results have been deferred to the Appendices A and B.

1.4 Notation

Below, we collect some notation that is used throughout the manuscript. This notation is mostly standard; hence, this section can be skipped and only be referred to when a symbol is unclear.

We denote by $\mathbb{N} = \{1, 2, ...\}$ the set of all *natural numbers* and define, for $k \in \mathbb{N}$, the set $\mathbb{N}_{\geq k} := \{n \in \mathbb{N} \colon n \geq k\}$. For $d_1, d_2 \in \mathbb{N}$ we denote by $\mathrm{Id}_{\mathbb{R}^{d_1}}$ the *identity* on $\mathbb{R}^{d_1}$ and by $0_{\mathbb{R}^{d_1 \times d_2}}$ we denote the map from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$ that vanishes everywhere. We denote by $0_{\mathbb{R}^{d_1}}$ the *zero vector* in $\mathbb{R}^{d_1}$. On $\mathbb{R}^{d_1 \times d_2}$ we denote by $\|\cdot\|$ the *Euclidean norm* and by $\|\cdot\|_\infty$ the *maximum norm*. The *number of nonzero entries* of a matrix or vector $A \in \mathbb{R}^{d_1 \times d_2}$ is counted by $\|\cdot\|_0$, where $\|A\|_0 := |\{(i, j) : A_{i,j} \neq 0\}|$.

If $d_1, d_2, d_3 \in \mathbb{N}$, and $A \in \mathbb{R}^{d_1 \times d_2}$, $B \in \mathbb{R}^{d_1 \times d_3}$, then we use the *block matrix notation* and write for the horizontal concatenation of A and B

$$\begin{bmatrix} A & B \end{bmatrix} \in \mathbb{R}^{d_1, d_2+d_3} \quad \text{or} \quad \begin{bmatrix} A \,\big|\, B \end{bmatrix} \in \mathbb{R}^{d_1, d_2+d_3},$$

where the second notation is used if a stronger delineation between different blocks is appropriate. A similar notation is used for the vertical concatenation of $A \in \mathbb{R}^{d_1 \times d_2}$ and $B \in \mathbb{R}^{d_3 \times d_2}$.

For $d_1, d_2 \in \mathbb{N}$, and $\Omega \subset \mathbb{R}^{d_1}$, we denote by $L^p(\Omega, \mathbb{R}^{d_2})$, $p \in [1, \infty]$ the $\mathbb{R}^{d_2}$-*valued Lebesgue spaces*, where we set $L^p(\Omega) := L^p(\Omega, \mathbb{R})$. For $k \in \mathbb{N}$, we denote by $W^{k,\infty}(\Omega)$, the space of *k-times weakly differentiable functions that have all derivatives of order at most k in $L^\infty(\Omega)$*. The space $W^{k,\infty}_{\mathrm{loc}}(\Omega)$ consists of functions such that their restriction to every compact $K \subset \Omega$ is in $W^{k,\infty}(K)$. By $C^k(\Omega, \mathbb{R}^{d_2})$, we denote the set of *$k$-times continuously differentiable functions* mapping from Ω to $\mathbb{R}^{d_2}$, where we set $C^k(\Omega) := C^k(\Omega, \mathbb{R})$. By $C^k_c(\Omega)$, we denote all functions in $C^k(\Omega)$ that have compact support.

For a Lipschitz continuous function $f : \mathbb{R}^{d_1} \mapsto \mathbb{R}^{d_2}$, we denote:

$$\mathrm{Lip}_f := \sup_{x \neq y} \frac{\|f(x) - f(y)\|}{\|x - y\|}.$$

Let $a > 0$, then we say for two functions $f : (0, a) \to [0, \infty)$ and $g : (0, a) \to [0, \infty)$ that $f(\varepsilon)$ is in $\mathcal{O}(g(\varepsilon))$ for $\varepsilon \to 0$ if there exists $0 < \delta < a$ and $C > 0$ such that $f(\varepsilon) \leq C g(\varepsilon)$ for all $\varepsilon \in (0, \delta)$.

2 Neural networks

In this section, we define neural networks and then recall a couple of operations on these objects that will be used frequently in the sequel. In the definition of neural networks, we distinguish between a neural network as a set of weights and an associated function that we call the realization of the neural network. This formal approach was introduced in [48], but we recall here a slightly different formulation of [28] for neural networks that allow so-called skip connections.

Definition 2.1 Let $d, L \in \mathbb{N}$. A *neural network (NN) Φ with input dimension d and L layers* is a sequence of matrix-vector tuples

$$\Phi = ((A_1, b_1), (A_2, b_2), \ldots, (A_L, b_L)),$$

where, for $N_0 = d$ and $N_1, \ldots, N_L \in \mathbb{N}$, each A_ℓ is an $N_\ell \times \sum_{k=0}^{\ell-1} N_k$ matrix, and $b_\ell \in \mathbb{R}^{N_\ell}$.

Let $\varrho : \mathbb{R} \to \mathbb{R}$ be the ReLU, i.e., $\varrho(x) = \max\{0, x\}$ and let Φ be a NN as above. Then, we define the associated *realization of* Φ as the map $\mathrm{R}(\Phi) : \mathbb{R}^d \to \mathbb{R}^{N_L}$ such that

$$\mathrm{R}(\Phi)(x) = x_L,$$

where x_L results from the following scheme:

$$\begin{aligned} x_0 &:= x, \\ x_\ell &:= \varrho\left(A_\ell \left[\, x_0^T \middle| \ldots \middle| x_{\ell-1}^T \,\right]^T + b_l\right), \quad \text{for } \ell = 1, \ldots L-1, \\ x_L &:= A_L \left[\, x_0^T \middle| \ldots \middle| x_{L-1}^T \,\right]^T + b_L. \end{aligned}$$

Here, ϱ acts componentwise, i.e., $\varrho(y) = [\varrho(y^1), \ldots, \varrho(y^m)]$ for $y = [y^1, \ldots, y^m] \in \mathbb{R}^m$. We sometimes write A_ℓ in block-matrix form as

$$A_\ell = \left[\, A_{\ell,0} \middle| \ldots \middle| A_{\ell,\ell-1} \,\right],$$

where $A_{\ell,k}$ is an $N_\ell \times N_k$ matrix for $k = 0, \ldots, \ell - 1$ and $\ell = 1, \ldots, L$. Then

$$\begin{aligned} x_\ell &= \varrho\left(A_{\ell,0}x_0 + \ldots + A_{\ell,\ell-1}x_{\ell-1} + b_\ell\right), \quad \text{for } \ell = 1, \ldots L-1, \\ x_L &= A_{L,0}x_0 + \ldots + A_{L,L-1}x_{L-1} + b_L. \end{aligned}$$

We call $N(\Phi) := d + \sum_{j=1}^L N_j$ the *number of neurons* of the NN Φ, $L = L(\Phi)$ the *number of layers*, and $W(\Phi) := \sum_{j=1}^L (\|A_j\|_0 + \|b_j\|_0)$ is called the *number of weights* of Φ. Moreover, we refer to N_L as *output dimension* of Φ.

2.1 Standard operations on neural networks

We collect four standard operations that can be performed with NNs below. First, we can concatenate two NNs Φ^1, Φ^2 in such a way that the realization of the concatenation is a composition of the individual realizations of Φ^1, Φ^2.

Proposition 2.2 ([28, Remark 2.8]) *Let Φ^1, Φ^2 be two NNs such that the input dimension d of Φ^1 is equal to the output dimension of Φ^2. Then, there exists a NN $\Phi^1 \odot \Phi^2$ such that*

- $L\left(\Phi^1 \odot \Phi^2\right) = L\left(\Phi^1\right) + L\left(\Phi^2\right)$,
- $W\left(\Phi^1 \odot \Phi^2\right) \le 2W\left(\Phi^1\right) + 2W\left(\Phi^2\right)$,
- $\mathrm{R}\left(\Phi^1 \odot \Phi^2\right)(x) = \mathrm{R}\left(\Phi^1\right) \circ \mathrm{R}\left(\Phi^2\right)(x)$ *for all* $x \in \mathbb{R}^d$.

We call $\Phi^1 \odot \Phi^2$ the sparse concatenation of Φ^1 and Φ^2.

An additional operation that is frequently applied to NNs in the sequel is that of parallelization. This procedure puts NNs in parallel such that the output of the realization is a vector containing the outputs of the original NNs.

Proposition 2.3 ([28, Remark 2.10]) *Let $n, d \in \mathbb{N}$ and, for $i = 1, \ldots, n$, let Φ^i be a NN with d-dimensional input and $L_i \in \mathbb{N}$ layers. Then there exists a NN* $\mathrm{P}(\Phi^1, \ldots, \Phi^n)$ *with d-dimensional input such that*

- $L\left(\mathrm{P}\left(\Phi^1, \ldots, \Phi^n\right)\right) = \max\{L_1, \ldots, L_n\}$,
- $W\left(\mathrm{P}\left(\Phi^1, \ldots, \Phi^n\right)\right) = \sum_{i=1}^n W\left(\Phi^i\right)$,
- $\mathrm{R}\left(\mathrm{P}\left(\Phi^1, \ldots, \Phi^n\right)\right)(x) = \left(\mathrm{R}\left(\Phi^1\right)(x), \ldots, \mathrm{R}\left(\Phi^n\right)(x)\right)$ *for all* $x \in \mathbb{R}^d$.

We call $\mathrm{P}(\Phi^1, \ldots, \Phi^n)$ *the parallelization of* $\Phi^1, \ldots, \Phi^n$.

We will occasionally need to construct NNs the realization of which is the sum of functions that we had approximated beforehand by realizations of NNs. In this situation, the following operation that emulates a sum of NNs is convenient.

Proposition 2.4 (Sum of NNs) *Let* $d \in \mathbb{N}$*, and* Φ^1, Φ^2 *be two NNs with* d*-dimensional input and one-dimensional output. Then there exists a NN* $\Phi^1 \oplus \Phi^2$ *with* d*-dimensional input such that*

- $L\left(\Phi^1 \oplus \Phi^2\right) = \max\{L(\Phi^1), L(\Phi^2)\}$,
- $W\left(\Phi^1 \oplus \Phi^2\right) = W(\Phi^1) + W(\Phi^2)$,
- $\mathrm{R}\left(\Phi^1 \oplus \Phi^2\right)(x) = \mathrm{R}(\Phi^1)(x) + \mathrm{R}(\Phi^2)(x)$ *for all* $x \in \mathbb{R}^d$.

We call $\Phi^1 \oplus \Phi^2$ *the sum of* Φ^1 *and* Φ^2.

Proof Let

$$((A_1, b_1), (A_2, b_2), \ldots, (A_L, b_L)) := \mathrm{P}(\Phi^1, \Phi^2).$$

Then we set

$$\widetilde{A}_L := \left[\, 1 \;\; 1 \,\right] A_L \quad \text{and}\; \widetilde{b}_L := \left[\, 1 \;\; 1 \,\right] b_L.$$

Clearly, $\|\widetilde{A}_L\|_{\ell_0} \leq \|A_L\|_{\ell_0}$ and $\|\tilde{b}_L\|_{\ell_0} \leq \|b_L\|_{\ell_0}$. We define

$$\Phi^1 \oplus \Phi^2 := \left((A_1, b_1), (A_2, b_2), \ldots, (A_{L-1}, b_{L-1}), \left(\widetilde{A}_L, \tilde{b}_L\right)\right).$$

Per construction,

$$\mathrm{R}\left(\Phi^1 \oplus \Phi^2\right)(x) = \left[\, 1 \;\; 1 \,\right] \begin{pmatrix} \mathrm{R}\left(\Phi^1\right)(x) \\ \mathrm{R}(\Phi^2)(x) \end{pmatrix} = \mathrm{R}\left(\Phi^1\right)(x) + \mathrm{R}\left(\Phi^2\right)(x) \quad \text{for every } x \in \mathbb{R}^d. \qquad \square$$

Finally, we can construct a NN that represents the multiplication of two NNs Φ^1 and Φ^2 in the sense that its realization is close to the multiplication of the realizations of Φ^1 and Φ^2. In contrast to the previous operations, this emulation of the multiplication is not exact but requires a parameter $\varepsilon > 0$ describing how accurately the multiplication is implemented.

Proposition 2.5 (Multiplication of NNs) *Let* Φ^1, Φ^2 *be NNs with input dimensions* d_1 *and* d_2 *and output dimension 1. Then, for every* $\varepsilon \in (0, 1)$*, there exists a NN* $\Phi^1 \otimes^\varepsilon \Phi^2$ *such that, for a universal constant* $c_1 > 0$ *and for* $c_2 = c_2(\|\mathrm{R}(\Phi^1)\|_{L^\infty}, \|\mathrm{R}(\Phi^2)\|_{L^\infty}) > 0$*, there holds*

- $L\left(\Phi^1 \otimes^\varepsilon \Phi^2\right) \leq \max\{L\left(\Phi^1\right), L\left(\Phi^2\right)\} + c_1 \ln(1/\varepsilon) + c_2$,
- $W\left(\Phi^1 \otimes^\varepsilon \Phi^2\right) \leq c_1 \ln(1/\varepsilon) + c_2 + 2W\left(\Phi^1\right) + 2W(\Phi^2)$,
- $\left\|\mathrm{R}\left(\Phi^1 \otimes^\varepsilon \Phi^2\right) - \mathrm{R}\left(\Phi^1\right)\mathrm{R}\left(\Phi^2\right)\right\|_{L^\infty} \leq \varepsilon$.

Proof By [57, Proposition 3], there exists, for every $\varepsilon \in (0, 1)$ and $M \in \mathbb{N}$, a NN $\times^{\varepsilon,M}$ with two-dimensional input and one-dimensional output satisfying

$$\left|\mathrm{R}\left(\times^{\varepsilon,M}\right)(x, y) - xy\right| \leq \varepsilon,$$

for all $x, y \in [-M, M]$. Moreover,

$$W\left(\times^{\varepsilon,M}\right) \leq c_2 - c_1 \ln(\varepsilon), \quad L\left(\times^{\varepsilon,M}\right) \leq c_2 - c_1 \ln(\varepsilon)$$

for a universal constant c_1 and $c_2 = c_2(M)$.

We set $\widetilde{M} := \max\{\|\mathrm{R}(\Phi^1)\|_{L^\infty}, \|\mathrm{R}(\Phi^2)\|_{L^\infty}\}$ and define

$$\Phi^1 \otimes^{\varepsilon} \Phi^2 := \times^{\varepsilon,\widetilde{M}} \odot \mathrm{P}(\Phi_1, \Phi_2).$$

The result now follows from Propositions 2.2 and 2.3. □

2.2 Approximation of smooth functions

In addition to the operations on NNs described in the previous section, we will frequently invoke the following standard approximation result of smooth functions by realisations of NNs.

Theorem 2.6 ([57, Theorem 1]) *Let $k, d \in \mathbb{N}$ and*

$$F_{k,d} := \left\{f \in W^{k,\infty}\left([0, 1]^d\right) : \|f\|_{W^{k,\infty}([0,1]^d)} \leq 1\right\}.$$

Then there exists $c = c(k, d) > 0$ such that, for every $f \in F_{k,d}$ and every $\varepsilon \in (0, 1)$, there exists a NN $\Phi^{f,\varepsilon}$ with d-dimensional input such that,

- $L(\Phi^{f,\varepsilon}) \leq c \cdot (\ln(1/\varepsilon) + 1)$,
- $W(\Phi^{f,\varepsilon}) \leq c\, \varepsilon^{-d/k} \cdot (\ln(1/\varepsilon) + 1)$,
- $\|f - \mathrm{R}(\Phi^{f,\varepsilon})\|_{L^\infty} < \varepsilon$.

Remark 2.7 (i) The norm we use for $W^{k,\infty}([0, 1]^d)$ is

$$\|f\|_{W^{k,\infty}([0,1]^d)} := \max_{\alpha:|\alpha|\leq k} \operatorname*{ess\,sup}_{x\in[0,1]^d} |D^\alpha f(x)|.$$

(ii) The space $W^{k,\infty}([0, 1]^d)$ can be identified with the set of $k-1$-times continuously differentiable functions all derivatives of order $k-1$ of which are Lipschitz continuous.

(iii) If we consider the ball with radius R in $W^{k,\infty}([0, 1]^d)$, i.e.,

$$F^R_{k,d} := \left\{f \in W^{k,\infty}([0, 1]^d) : \|f\|_{W^{k,\infty}([0,1]^d)} \leq R\right\}$$

instead of the unit ball $F_{k,d}$ then the constant c from Theorem 2.6 also depends on R. However, the asymptotic behavior with respect to ε remains unchanged. The same change holds if we consider the space $W^{k,\infty}([0, R]^d)$ instead of $W^{k,\infty}([0, 1]^d)$.

3 Linear transport equations

In this section, we introduce the Cauchy problem for the parametric linear transport equation, state the most important existence results for several types of linear transport equations, and provide expressions for their solutions. Here, we mainly follow the French book [24]. An English translation of this source can be found in the lecture notes [25]. For more information on linear transport equations, see also [2, 3, 10].

Definition 3.1 The Cauchy problem of the parametric linear transport equation is given by

$$\begin{cases} \partial_t u(t,x,\eta) + V(t,x,\eta)\cdot\nabla_x u(t,x,\eta) = 0, & \text{(3.1a)}\\ u(0,x,\eta) = u_0(x), & \text{(3.1b)} \end{cases}$$

where $t \in [0,T]$, $x \in \mathbb{R}^n$, and $\eta \in [0,1]^D$ for some $n, D \in \mathbb{N}$, and $T > 0$. The vector field $V \in C^k([0,T]\times\mathbb{R}^n\times[0,1]^D;\mathbb{R}^n)$ and the initial condition $u_0 \in C^s(\mathbb{R}^n;\mathbb{R})$ are given with $s, k \in \mathbb{N}$.

It is well known that linear transport equations can be solved via the method of characteristics [12, 22, 35]. The idea of this method is to consider characteristic curves that are defined so that the solution u of (3.1) is constant along these curves. Then, the solution at a point (t,x,η) equals the initial data evaluated at the origin of this curve.

Definition 3.2 The *characteristic curve* of the transport operator $\partial_t + V(t,x,\eta)\cdot\nabla_x$ passing through x at time $s=t$ is given by the set $\{(s,\gamma(s)) : s \in [0,T]\}$, where γ is the solution of the *characteristic system of ordinary differential equations*

$$\begin{cases} \dot\gamma(s) = V(s,\gamma(s),\eta), & \text{(3.2a)}\\ \gamma(t) = x. & \text{(3.2b)} \end{cases}$$

Let us briefly show why the solution u of (3.1) does not change along characteristic curves. Considering the case where $V(t,x,\eta) \equiv v$ for a $v \in \mathbb{R}^n$ and dropping the η-dependency, we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}u(t,\gamma(t)) &= \partial_t u(t,\gamma(t)) + \nabla_x u(t,\gamma(t))\cdot\dot\gamma(t)\\ &= \partial_t u(t,\gamma(t)) + \nabla_x u(t,\gamma(t))\cdot v\\ &= (\partial_t + V(t,x)\cdot\nabla_x)u(t,\gamma(t)) = 0, \end{aligned}$$

where the last equality is due to (3.1).

To make the method of characteristics work, we have to ensure that the characteristic curves are diffeomorphisms and that there exists a global solution of system (3.2). Therefore, we make the following assumptions on the vector field V: Let $n, D \in \mathbb{N}$, and $T > 0$.

(H1) For some $k \in \mathbb{N}$, there holds $V \in C^k([0,T]\times\mathbb{R}^n\times[0,1]^D;\mathbb{R}^n)$.

(H2) There exists a $C > 0$ s.t.

$$|V(t, x, \eta)| \leq C(1 + |x|) \quad \text{for all } (t, x, \eta) \in [0, T] \times \mathbb{R}^n \times [0, 1]^D.$$

The following theorem states that these assumptions lead to global existence of the characteristic curves and characterizes their regularity.

Theorem 3.3 ([25, Theorem 2.2.2]) *Let V satisfy (H1) and (H2) with $n, D, k \in \mathbb{N}$, and $T > 0$. Then, for all $(t, x, \eta) \in [0, T] \times \mathbb{R}^n \times [0, 1]^D$, the system of (3.2) has a unique solution $\gamma \in C^k([0, T])$. Furthermore, the map X defined by*

$$X(s, t, x, \eta) := \gamma(s)$$

is in $C^k([0, T] \times [0, T] \times \mathbb{R}^n \times [0, 1]^D)$.

Proof The proof presented in [25] can directly be extended to the parametric case. Moreover, the differentiability with respect to η is a standard result: compare [29, Corollary 4.1]. □

Remark 3.4 The previous theorem implies that X is bounded on $[0, T] \times [0, T] \times K \times [0, 1]^D$ for every compact domain $K \subset \mathbb{R}^n$. Therefore, the final bound in our main result depends additionally on K and the velocity field V. An explicit bound for $\|X\|_{C^k}$ in terms of the given data seems to be very technical to derive. In Appendix A, we compute precise bounds for $k = 0$ and $k = 1$, which are explicitly used in in the proof of Theorem 4.3 and Theorem 4.8. The argument there suggests that a bound on $\|X\|_{C^k}$ in terms of $|K|$ and $\|V\|_{C^k}$ could likely be established, however with considerable technical effort.

3.1 Solutions of the standard linear transport equation

The next theorem states the existence of a solution for the linear transport equation of (3.1). Furthermore, the theorem establishes that the solution has a compositional structure resulting from composing the initial data with the solution of the characteristic system of ODEs starting at (t, x, η) evaluated at $s = 0$.

Theorem 3.5 ([25, Theorem 2.2.4]) *Let V satisfy the assumptions (H1) and (H2) with $n, D, k \in \mathbb{N}$, and $T > 0$. Furthermore, let $u_0 \in C^s(\mathbb{R}^n)$, $s \in \mathbb{N}$. Then, the Cauchy problem for the parametric linear transport equation of (3.1) has a unique solution $u \in C^{\min\{s,k\}}([0, T] \times \mathbb{R}^n \times [0, 1]^D)$ which is given by*

$$u(t, x, \eta) = u_0(X(0, t, x, \eta)).$$

For initial conditions that are not differentiable, it makes sense to introduce a weak notion of a solution. The following definition and proposition were taken from [18]. A proof for the simplified case, where $\mathrm{div}_x V = 0$ can be found in [42, Theorem 3.12].

Definition 3.6 ([18]) Let V satisfy the assumptions *(H1)* and *(H2)* with $n, D, k \in \mathbb{N}$, and $T > 0$. Furthermore, let $u_0 \in L^\infty(\mathbb{R}^n)$. A weak solution to (3.1) is a function $u \in L^\infty([0, T] \times \mathbb{R}^n \times [0, 1]^D)$ which satisfies the *weak formulation*

$$\int_0^T \int_{\mathbb{R}^n} u(t, x, \eta) \left[\partial_t \varphi + V(t, x, \eta) \cdot \nabla_x \varphi(t, x) + \operatorname{div}_x V(t, x, \eta) \varphi(t, x)\right] \mathrm{d}x \, \mathrm{d}t \\ + \int_{\mathbb{R}^n} u_0(x) \varphi(0, x) \, \mathrm{d}x = 0 \tag{3.3}$$

for all $\varphi \in C_c^1([0, T) \times \mathbb{R}^n)$ and all $\eta \in [0, 1]^D$.

As for strong solutions of the transport equation, the solution of the weak formulation is given by a composition of the initial condition with a flow along the characteristic curves.

Proposition 3.7 ([18]) *Let V satisfy the assumptions (H1) and (H2) with $n, D, k \in \mathbb{N}$, and $T > 0$. Furthermore, let $u_0 \in L^\infty(\mathbb{R}^n)$. Then there exists a global weak solution $u \in L^\infty([0, T] \times \mathbb{R}^n \times [0, 1]^D)$ to (3.1) which is given by*

$$u(t, x, \eta) = u_0(X(0, t, x, \eta)).$$

3.2 Solutions of extensions of the parametric linear transport equation

In Section 4, we extend our main result to linear transport equations that include source terms and are formulated in conservative form. The following two propositions present the corresponding existence results and the form of the solutions for problems with source terms and in conservative form.

Proposition 3.8 ([42, Theorem 3.9]) *Let V satisfy assumptions (H1) and (H2) with $n, D, k \in \mathbb{N}$, and $T > 0$. Furthermore, let $u_0 \in C^s(\mathbb{R}^n)$ and $f \in C^{s'}([0, T] \times \mathbb{R}^n \times [0, 1]^D)$, where $s, s' \in \mathbb{N}$. Then the Cauchy problem for the non-homogeneous parametric linear transport equation*

$$\begin{cases} \partial_t u(t, x, \eta) + V(t, x, \eta) \cdot \nabla_x u(t, x, \eta) = f(t, x, \eta), & (3.4a) \\ u(0, x, \eta) = u_0(x), & (3.4b) \end{cases}$$

has a unique solution $u \in C^{\min\{s,s',k\}}([0, T] \times \mathbb{R}^n \times [0, 1]^D)$ which is given by

$$u(t, x, \eta) = u_0(X(0, t, x, \eta)) + \int_0^t f(s, X(s, t, x, \eta), \eta) \, \mathrm{d}s. \tag{3.5}$$

Remark 3.9 Similar to Proposition 3.7, one can prove the existence and uniqueness of a weak solution with a source term $f \in C^0([0, T] \times \mathbb{R}^n \times [0, 1]^D)$. In this case, the associated weak formulation is given by (3.3) after replacing the right-hand side by $\int_0^T \int_{\mathbb{R}^n} f(t, x, \eta) \varphi(t, x, \eta) \, \mathrm{d}x \, \mathrm{d}t$. The weak solution $u \in L^\infty([0, T] \times \mathbb{R}^n \times [0, 1]^D)$ of that problem is still given by (3.5). Here, one only needs to assume that u_0 is continuous, [42, Remark 3.13] or [18].

Proposition 3.10 ([25, Theorem 2.3.6]) *Let V satisfy assumptions (H1) and (H2) with $n, D, k \in \mathbb{N}$, and $T > 0$. Furthermore, let $u_0 \in C^s(\mathbb{R}^n)$, $s \in \mathbb{N}$. Then the Cauchy problem for the conservative parametric linear transport equation:*

$$\begin{cases} \partial_t u(t,x,\eta) + \operatorname{div}_x(V(t,x,\eta)u(t,x,\eta)) = 0, & \text{(3.6a)} \\ u(0,x,\eta) = u_0(x), & \text{(3.6b)} \end{cases}$$

has a unique solution $u \in C^{\min\{s,k\}}([0,T] \times \mathbb{R}^n \times [0,1]^D)$ which is given by

$$u(t,x,\eta) = u_0(X(0,t,x,\eta))J(0,t,x,\eta) \tag{3.7}$$

with

$$J(s,t,x,\eta) = \det(D_x X(s,t,x,\eta)).$$

Remark 3.11 Again, one can show that there exists a unique weak solution for the conservative formulation which is given by (3.7). See [25, Section 2.3] for more information about this problem.

Remark 3.12 The conservative form (3.6) simplifies to the original linear transport (3.1) if $\operatorname{div}_x V = 0$.

4 DNN approximation of solutions of linear transport equations

Theorem 3.5 and Propositions 3.8 and 3.10 suggest that the solutions of parametric linear transport equations are of a compositional form, where the initial condition is composed with a flow along characteristic curves. Since realizations of NNs are naturally of compositional structure, it is therefore conceivable that the form of the solutions of linear transport equations can be efficiently resolved by NNs. Indeed, based on this observation we present, for each of the cases discussed in Section 3, an approximation result for the solution of the associated parametric linear transport equations by NNs.

4.1 Standard linear transport equations

We start by presenting an approximation result for the solutions of standard linear transport equations as described in Theorem 3.5. We will assume that the initial condition can be approximated reasonably well by NNs. For this, we use the following definition:

Definition 4.1 Let $n \in \mathbb{N}$ and $r > 0$, a function $f \in L^\infty(\mathbb{R}^n)$ is *r-approximable by NNs* if, for every compact set $K \subset \mathbb{R}^n$, there exists a constant $c = c(K, r, f) > 0$ such that for every $\epsilon \in (0,1)$ there exists a NN $\Phi^{f,\varepsilon}$ such that

- $L\left(\Phi^{f,\varepsilon}\right) \leq c \cdot (\ln(1/\varepsilon) + 1)$,
- $W\left(\Phi^{f,\varepsilon}\right) \leq c\, \varepsilon^{-1/r}(\ln(1/\epsilon) + 1)$,

- $\|f - \mathrm{R}(\Phi)\|_{L^\infty(K)} \leq \varepsilon.$

Remark 4.2 By Theorem 2.6, every function $f \in C^s(\mathbb{R}^n)$ is r approximable for $r = s/n$.

We now present the main theorems of this section for the strong and weak formulations of standard linear transport equations below. Afterward, in Remark 4.2, we discuss to what extent the resulting approximation rates improve upon a direct application of Theorem 2.6 to the solution u of a linear transport equation. We present the proofs of the theorems at the end of this subsection.

Theorem 4.3 *Let V satisfy assumptions (H1) and (H2) for $k, n, D \in \mathbb{N}$, and $T > 0$. Further let, for $r > 0$, $u_0 \in C^1(\mathbb{R}^n)$ be r-approximable by NNs. Let $u \in C^1([0,T] \times \mathbb{R}^n \times [0,1]^D)$ denote the unique solution of the Cauchy problem for the parametric linear transport equation*

$$\begin{cases} \partial_t u(t,x,\eta) + V(t,x,\eta) \cdot \nabla_x u(t,x,\eta) = 0, \\ u(0,x,\eta) = u_0(x). \end{cases}$$

Then, for every $\varepsilon \in (0,1)$ and every compact subset $K \subset \mathbb{R}^n$, there exists a NN $\Phi^{\overline{u},\varepsilon}$ with d-dimensional input, where $d := 1 + n + D$, such that for the restriction $\overline{u} := u\big|_{[0,T]\times K\times[0,1]^D}$ there holds that, for $c = c(n,r,d,k,K,T,V,u_0) > 0$,

(i) $L\left(\Phi^{\overline{u},\varepsilon}\right) \leq c \cdot (\ln(1/\varepsilon) + 1)$,
(ii) $W(\Phi^{\overline{u},\varepsilon}) \leq c \cdot \left(\varepsilon^{-1/r} + \varepsilon^{-d/k}\right) \cdot (\ln(1/\varepsilon) + 1)$,
(iii) $\left\|\overline{u} - \mathrm{R}\left(\Phi^{\overline{u},\varepsilon}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)} < \varepsilon$,

In Theorem 4.3 above, the initial condition is required to be continuously differentiable. In the following result, we extend Theorem 4.3 to initial conditions that are Lipschitz continuous only. To handle initial conditions that are not continuously differentiable, we have to consider weak solutions as described in Definition 3.6.

Theorem 4.4 *Let V satisfy assumptions (H1) and (H2) for $k, n, D \in \mathbb{N}$, and $T > 0$. Further let, for $r > 0$, $u_0 \in W^{1,\infty}_{\mathrm{loc}}$ be r-approximable by NNs. Let $u(t,x,\eta) = u_0(X(0,t,x,\eta))$ denote the weak solution of the Cauchy problem for the parametric linear transport equation of* (3.1) *according to* Proposition 3.7.

Then, for every $\varepsilon \in (0,1)$ and every compact subset $K \subset \mathbb{R}^n$, there exists a NN $\Phi^{\overline{u},\varepsilon}$ with d-dimensional input, where $d := 1 + n + D$, such that for the restriction $\overline{u} := u\big|_{[0,T]\times K\times[0,1]^D}$ there holds that

(i) $L\left(\Phi^{\overline{u},\varepsilon}\right) \leq c \cdot (\ln(1/\varepsilon) + 1)$,
(ii) $W\left(\Phi^{\overline{u},\varepsilon}\right) \leq c \cdot \left(\varepsilon^{-1/r} + \varepsilon^{-d/k}\right) \cdot (\ln(1/\varepsilon) + 1)$,
(iii) $\left\|\overline{u} - \mathrm{R}\left(\Phi^{\overline{u},\varepsilon}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)} < \varepsilon$,

for $c = c(n,r,d,k,K,T,V,u_0) > 0$.

Proof of Theorem 4.3 Recall that, by Theorem 3.5, the unique solution $u \in C^1([0, T] \times \mathbb{R}^n \times [0,1]^D)$ of (3.1) is given by

$$u(t,x,\eta) = u_0(X(0,t,x,\eta)).$$

Moreover, $X \in C^k([0,T] \times [0,T] \times \mathbb{R}^n \times [0,1]^D)$ by Theorem 3.3. Therefore, $\widetilde{X} := X(0,\cdot,\cdot,\cdot) \in C^k([0,T] \times \mathbb{R}^n \times [0,1]^D)$.

The idea of the proof is to first approximate the functions u_0 and $\widetilde{X}$ separately by realizations of NNs using Theorem 2.6 and then to concatenate these NNs by Proposition 2.2. To apply Theorem 2.6, we restrict the function $\widetilde{X}$ to $U := [0,T] \times K \times [0,1]^D$ and u_0 to $B_G(0)$ where $B_G(0) \subset \mathbb{R}^n$ denotes the ball of radius G around 0 with $G = (|K| + CT)\exp(CT)$ from (A.1). Then, Definition 4.1 and Theorem 2.6 imply that there exist NNs Φ^{u_0,δ_1} and $\Phi^{\widetilde{X},\delta_2}$ with n and d dimensional input dimension, respectively, such that for $\delta_1 := \varepsilon/2$ and $\delta_2 := \varepsilon/(2\,\mathrm{Lip}_{u_0})$ there holds

$$\left\|u_0 - \mathrm{R}\left(\Phi^{u_0,\delta_1}\right)\right\|_{L^\infty(B_G(0))} < \delta_1 \qquad \text{and} \qquad \left\|\widetilde{X} - \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right\|_{L^\infty(U)} < \delta_2.$$

Invoking the triangle inequality, we conclude for the concatenated network $\Phi^{u,\varepsilon} := \Phi^{u_0,\delta_1} \odot \Phi^{\widetilde{X},\delta_2}$ that

$$\begin{aligned}
\left\|u - \mathrm{R}\left(\Phi^{u,\varepsilon}\right)\right\|_{L^\infty(U)} &= \left\|u_0 \circ \widetilde{X} - \mathrm{R}\left(\Phi^{u_0,\delta_1} \odot \Phi^{\widetilde{X},\delta_2}\right)\right\|_{L^\infty(U)} \\
&= \left\|u_0 \circ \widetilde{X} - \mathrm{R}\left(\Phi^{u_0,\delta_1}\right) \circ \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right\|_{L^\infty(U)} \\
&\leq \left\|u_0 \circ \widetilde{X} - u_0 \circ \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right\|_{L^\infty(U)} \\
&\quad + \left\|u_0 \circ \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right) - \mathrm{R}\left(\Phi^{u_0,\delta_1}\right) \circ \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right\|_{L^\infty(U)} \\
&\leq \mathrm{Lip}_{u_0} \left\|\widetilde{X} - \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right\|_{L^\infty(U)} + \left\|u_0 - \mathrm{R}\left(\Phi^{u_0,\delta_1}\right)\right\|_{L^\infty(B_G(0))} \\
&\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

Additionally, we compute the number of weights of $\Phi^{u,\varepsilon}$ using Proposition 2.2, Definition 4.1, Theorem 2.6, Remark 2.7, and Remark 3.4 as

$$W\left(\Phi^{u,\varepsilon}\right) \leq 2 \cdot \left(W\left(\Phi^{u_0,\delta_1}\right) + W\left(\Phi^{\widetilde{X},\delta_2}\right)\right) \leq c \cdot \left(\varepsilon^{-1/r} + \varepsilon^{-d/k}\right) \cdot \left(\ln\left(1/\varepsilon\right) + 1\right)$$

with $c = c(n,r,d,k,K,T,V,u_0) > 0$. Moreover, by Proposition 2.2, Definition 4.1, Theorem 2.6, Remark 2.7, and Remark 3.4, we have that $L(\Phi^{u,\varepsilon}) \leq c \cdot (\ln(1/\varepsilon) + 1)$. □

Proof of Theorem 4.4 The proof is very similar to the proof of Theorem 4.3, but requires an application of Proposition 3.7 instead of Theorem 3.3. All further estimates are then analogous to those in the proof of Theorem 4.3 where only Lipschitz continuity of u_0 was used. □

4.2 Framework for efficient approximation

Theorems 4.3 and 4.4 describe achievable approximation rates for solutions of parametric transport equations by realizations of NNs. To assess if these rates describe an efficient way of approximating these functions they should be contrasted with a direct approximation, that is not based on the special structure of these functions. Morally, we expect our results to yield improved rates if the solution of a parametric transport equation u is a high-dimensional, non-smooth function, but splits into a high-dimensional smooth function and a low-dimensional rough function.

Below, we describe two situations where the aforementioned splitting yields through Theorems 4.3 and 4.4 highly efficient approximation rates, i.e., rates that are significantly better than those resulting from a direct approximation of u by Theorem 2.6. Afterward, we add a concrete example where the described situation appears.

The typical framework in which we expect to apply Theorems 4.3 and 4.4 above is that where V is substantially smoother than the initial condition u_0. Concretely, in the following two situations we have that the approximation rates resulting from an application of Theorem 4.3 or 4.4 are significantly better than those resulting from a direct approximation of u by Theorem 2.6.

- *The initial condition u_0 is smooth but not sufficiently smooth to set off the high dimension of the parameter space:* Assume that, in the notation of Theorems 4.3 and 4.4, $n \leq s \ll d \leq k$, and $u_0 \in C^s$. In this situation, the dimension of the parameter space is significantly larger than the dimension of the physical domain. The dependence of V on the parameters is, however, very regular.

  Then $u \in C^s([0, T] \times K \times [0, 1]^D)$ and a direct application of Theorem 2.6 would yield an approximating network with a complexity bound for the number of weights of the form $c \cdot ((\ln(1/\varepsilon) + 1)\varepsilon^{-d/s})$. On the other hand, Theorem 2.6 yields that u_0 is 1-approximable by NNs and hence Theorem 4.3 yields a complexity bound that is not worse than $c \cdot ((\ln(1/\varepsilon) + 1)\epsilon^{-1}$.
- *The initial condition u_0 is rough but can be efficiently represented by realisations of NNs:* Assume that $D \in \mathbb{N}$, in the notation of Theorems 4.3 and 4.4, $n \ll n+D$, and $k = d$. Moreover, assume that $u_0 \notin W^{2,\infty}_{\mathrm{loc}}$, but $u_0 \in W^{1,\infty}_{\mathrm{loc}}$ and u_0 can be very efficiently represented by the realisation of a NN. A typical example is that u_0 is a ramp function along a hyperplane or a piecewise affine function. In this case, u_0 is r-approximable for every $r \in \mathbb{R}$. Since in applications, u_0 is often known, it is conceivable that one can establish r-approximability of u_0 without using classical smoothness-based arguments.

  Then we have that $u \notin W^{2,\infty}([0, T] \times K \times [0, 1]^D)$ and therefore a direct application of Theorem 2.6 to approximate u would again yield a NN with a complexity bound for the number of weights of the form $c \cdot ((\ln(1/\varepsilon) + 1)\varepsilon^{-d})$. On the other hand, Theorem 4.4 yields a complexity bound of $c \cdot ((\ln(1/\varepsilon) + 1)\varepsilon^{-1})$.

A practical example of the above setting is, for instance, given by monoenergetic radiative transfer models. For further information and numerical methods for these models, refer to [14, 15]. A core constituent of radiative transfer models is described by

$$\begin{cases} \partial_t u(t,x,\eta) + \eta \cdot \nabla_x u(t,x,\eta) + a(t,x,\eta)u(x,t,\eta) = f(t,x,\eta), \\ u(0,x,\eta) = u_0(x), \\ u(t,x,\eta) = u_b^-(t,x,\eta), \qquad \text{for } (x,\eta) \in \partial\Omega^-, \end{cases}$$

with

$$\partial\Omega^- := \left\{(x,\eta) \in \partial\Omega \times [0,1]^D \,:\, \eta \cdot \nu < 0\right\}$$

where ν denotes the unit outward normal vector of $\partial\Omega$. This model includes a source term f, a cross section function a and inflow boundary conditions u_b^-. We derive the corresponding approximation results for this more general form of a transport equation in the next sections. The above formulation belongs to the setting of dominating transport problems for the common class of kinetic models that describe the propagation of particles in a collisional medium. The parameter η can, for example, describe a unit direction vector taken from the $(n-1)$-dimensional unit sphere. In this case, the model assumes that all particles have the same kinetic energy. For this problem, the velocity field $V(t,x,\eta) = \eta$ obviously fulfils the assumptions (H1) and (H2) with $k = \infty$. Therefore, an approximation independent of the dimension of η is possible for suitable initial data.

4.3 Non-vanishing source term

In the following, we extend Theorem 4.3 to non-vanishing source terms. We state our results for two different types of source terms. For V satisfying assumptions (H1) and (H2) with $k, n, D \in \mathbb{N}$, and $T > 0$, we assume that one of the following properties holds:

$$\text{(i) } f(t,x,\eta) = f_1(t,x) \in C^{s'}([0,T] \times \mathbb{R}^n), \qquad \text{for } s' := \lceil (n+1)k/d \rceil, \tag{4.1}$$

$$\text{(ii) } f(t,x,\eta) = f_2(t,x,\eta) \in C^{s'}([0,T] \times \mathbb{R}^n \times [0,1]^D), \qquad \text{for } s' := k. \tag{4.2}$$

In words, we assume high regularity of f if it depends on η while much less regularity of f is sufficient for the η-independent case.

Remark 4.5 In both cases, (4.1) and (4.2), Theorem 2.6 demonstrates that for every $\varepsilon \in (0,1)$ there exists a NN $\Phi^{f,\varepsilon}$ such that

- $L(\Phi^{f,\varepsilon}) \le c \cdot (\ln(1/\varepsilon) + 1)$,
- $W(\Phi^{f,\varepsilon}) \le c\, \varepsilon^{-d/k} \cdot (\ln(1/\varepsilon) + 1)$,
- $\|f - \mathrm{R}(\Phi^{f,\varepsilon})\|_{L^\infty} < \varepsilon$.

Based on the assumption on the source term, we next present an approximation result for solutions of non-homogeneous parametric linear transport equations.

Theorem 4.6 *Let V satisfy assumptions (H1) and (H2) for $k, n, D \in \mathbb{N}$, and $T > 0$. Further let, for $r > 0$, $u_0 \in C^1(\mathbb{R}^n)$ be r-approximable by NNs and let f and s' be* *as in* (4.1) *or* (4.2). *Let* $u \in C^1([0,T] \times \mathbb{R}^n \times [0,1]^D)$ *denote the unique solution of the Cauchy problem for the non-homogeneous parametric linear transport equation*

$$\begin{cases} \partial_t u(t,x,\eta) + V(t,x,\eta) \cdot \nabla_x u(t,x,\eta) = f(t,x,\eta), \\ u(0,x,\eta) = u_0(x). \end{cases}$$

Then, for every $\varepsilon \in (0,1)$ *and every compact subset* $K \subset \mathbb{R}^n$, *there exists a NN* $\Phi^{\overline{u},\varepsilon}$ *with* d*-dimensional input, where* $d := 1 + n + D$, *such that for the restriction* $\overline{u} := u\big|_{[0,T]\times K \times [0,1]^D}$ *there holds that*

(i) $L\left(\Phi^{\overline{u},\varepsilon}\right) \leq c \cdot (\ln(1/\varepsilon) + 1)$,
(ii) $W\left(\Phi^{\overline{u},\varepsilon}\right) \leq c \cdot \left(\varepsilon^{-1/r} + \varepsilon^{-(d+1)/k-1}\right) \cdot (\ln(1/\varepsilon) + 1)$,
(iii) $\left\|\overline{u} - \mathrm{R}\left(\Phi^{\overline{u},\varepsilon}\right)\right\|_{L^\infty([0,T]\times K \times [0,1]^D)} < \varepsilon$,

for $c = c(n, r, d, k, K, T, V, u_0, \|f\|_{C^{s'}}) > 0$.

Remark 4.7
- If $u_0 \in C^s(\mathbb{R}^n)$, then Remark 4.2 demonstrates that we can replace r in Theorem 4.6 by s/n and the constant c in Theorem 4.6 depends more specifically on $\|u_0\|_{C^s}$.
- As for Theorem 4.3, one can immediately generalize Theorem 4.6 to represent solutions of weak formulations via Remark 3.9.

Proof of Theorem 4.6 Recall that, by Proposition 3.8, the unique solution $u \in C^1([0,T] \times \mathbb{R}^n \times [0,1]^D)$ of (3.4) is given by

$$u(t,x,\eta) = u_0(X(0,t,x,\eta)) + \int_0^t f(\tau, X(\tau,t,x,\eta),\eta)\,\mathrm{d}\tau, \tag{4.3}$$

for all $(t,x,\eta) \in [0,T] \times \mathbb{R}^n \times [0,1]^D$.

The proof, therefore, proceeds as follows: First, via Proposition B.1, we construct a NN the realization of which approximates the antiderivative of f with respect to the first coordinate by a NN. Then, we construct a second NN via Theorem 4.3 the realisation of which approximates $u_0 \circ X$. Finally, Proposition 2.4 yields a NN such that the associated realization approximates (4.3).

Concretely, let $G = G(k,d,T,K,V) > 0$ be the upper bound for $\|X\|_{C^k}$. Definition 4.1 and Theorem 2.6 imply that there exist NNs Φ^{u_0,δ_1}, Φ^{X,δ_2} and Φ^{f,δ_3} with n, $d+1$, and d-dimensional input respectively such that, for

$$\delta_1 := \varepsilon/6, \quad \delta_2 := \varepsilon/(12\max\{\mathrm{Lip}_{u_0}, \mathrm{Lip}_f\}), \quad \delta_3 := \varepsilon/12, \tag{4.4}$$

we have that

$$\begin{aligned} &\left\|u_0 - \mathrm{R}\left(\Phi^{u_0,\delta_1}\right)\right\|_{L^\infty(B_G(0))} < \delta_1, \\ &\left\|X - \mathrm{R}\left(\Phi^{X,\delta_2}\right)\right\|_{L^\infty([0,T]\times[0,T]\times K\times[0,1]^D)} < \delta_2, \quad \text{and} \\ &\left\|f - \mathrm{R}\left(\Phi^{f,\delta_3}\right)\right\|_{L^\infty([0,T]\times B_G(0)\times[0,1]^D)} < \delta_3. \end{aligned}$$

Let $\Phi_\tau := \left(\left(\left[\,1\;0\;\ldots\;0\,\right], 0\right)\right)$ be a NN with one layer and input dimension $2 + n + D$. Moreover, let

$$\Phi_\eta := \left(\left(\left[0_{\mathbb{R}^{D\times(2+n)}}\middle|\mathrm{Id}_{\mathbb{R}^D}\right], 0_{\mathbb{R}^D}\right)\right).$$

We have that $\mathrm{R}(\Phi_\tau)(\tau, t, x, \eta) = \tau$ and $\mathrm{R}(\Phi_\eta)(\tau, t, x, \eta) = \eta$, for all $(\tau, t, x, \eta) \in [0, T]\times[0, T]\times K\times[0, 1]^D$. Setting $\Phi^{X,\delta_2,\mathrm{full}} := \mathrm{P}(\Phi_\tau, \Phi^{X,\delta_2}, \Phi_\eta)$, we now have that

$$\sup_{(\tau,t,x,\eta)\in[0,T]\times[0,T]\times K\times[0,1]^D} \left|(\tau, X(\tau, t, x, \eta), \eta) - \mathrm{R}\left(\Phi^{X,\delta_2,\mathrm{full}}\right)\right| \le \delta_2.$$

Denoting $f^X(\tau, t, x, \eta) := f(\tau, X(\tau, t, x, \eta), \eta)$, we have by the triangle inequality that

$$\begin{aligned}
&\sup_{(\tau,t,x,\eta)\in[0,T]\times[0,T]\times K\times[0,1]^D} \left|f^X(\tau, t, x, \eta) - \mathrm{R}\left(\Phi^{f,\delta_3} \odot \Phi^{X,\delta_2,\mathrm{full}}\right)(\tau, t, x, \eta)\right| \\
&\le \sup_{(\tau,t,x,\eta)\in[0,T]\times[0,T]\times K\times[0,1]^D} \left|f^X(\tau, t, x, \eta) - f \circ \mathrm{R}\left(\Phi^{X,\delta_2,\mathrm{full}}\right)(\tau, t, x, \eta)\right| \\
&\quad + \sup_{(\tau,t,x,\eta)\in[0,T]\times[0,T]\times K\times[0,1]^D} \left|f \circ \mathrm{R}\left(\Phi^{X,\delta_2,\mathrm{full}}\right)(\tau, t, x, \eta)\right. \\
&\quad \left. - \mathrm{R}\left(\Phi^{f,\delta_3} \odot \Phi^{X,\delta_2,\mathrm{full}}\right)(\tau, t, x, \eta)\right| \\
&\le \mathrm{Lip}_f \delta_2 + \delta_3.
\end{aligned} \tag{4.5}$$

Let $\widetilde{A}$ be the matrix satisfying $\widetilde{A}(t, x, \eta) = (t, t, x, \eta)$ for $(t, x, \eta) \in [0, T] \times K \times [0, 1]^D$. Then, for

$$N := \left\lceil \frac{15}{\varepsilon} \max\left\{T^2 \|f^X\|_{C^1}, 1 + \|f\|_{L^\infty([0,1]\times B_G(0)\times[0,1]^D)}\right\}\right\rceil, \tag{4.6}$$

we define with Proposition B.1

$$\Phi^{f,\mathrm{anti}} := \widetilde{I}_N\left(\Phi^{f,\delta_3} \odot \Phi^{X,\delta_2,\mathrm{full}}\right) \odot \left((\widetilde{A}, 0)\right).$$

We have the following estimate:

$$\begin{aligned}
&\left|\int_0^t f(\tau, X(\tau, t, x, \eta), \eta)\,\mathrm{d}\tau - \mathrm{R}\left(\Phi^{f,\mathrm{anti}}\right)(t, x, \eta)\right| \\
&\underset{\text{Prop. B.2}}{\le} \left|I_N(f^X)(t, t, x, \eta) - \mathrm{R}\left(\Phi^{f,\mathrm{anti}}\right)(t, x, \eta)\right| + \frac{2T^2\|f^X\|_{C^1}}{N} \\
&\le \left|I_N(f^X)(t, t, x, \eta) - \mathrm{R}\left(\widetilde{I}_N\left(\Phi^{f,\delta_3} \odot \Phi^{X,\delta_2,\mathrm{full}}\right)\right)(t, t, x, \eta)\right| + \frac{2T^2\|f^X\|_{C^1}}{N} \\
&\underset{\text{Rem. B.3}}{\le} \left|I_N\left(f^X\right)(t, t, x, \eta) - I_N\left(\mathrm{R}\left(\Phi^{f,\delta_3} \odot \Phi^{X,\delta_2,\mathrm{full}}\right)\right)(t, t, x, \eta)\right| \\
&\quad + \frac{2T^2\|f^X\|_{C^1}}{N} + \frac{3\left\|\mathrm{R}\left(\Phi^{f,\delta_3} \odot \Phi^{X,\delta_2,\mathrm{full}}\right)\right\|_{L^\infty}}{N} \\
&\le \left|I_N\left(f^X\right)(t, t, x, \eta) - I_N\left(\mathrm{R}\left(\Phi^{f,\delta_3} \odot \Phi^{X,\delta_2,\mathrm{full}}\right)\right)(t, t, x, \eta)\right| \\
&\quad + \frac{2T^2\|f^X\|_{C^1}}{N} + \frac{3\|\mathrm{R}\left(\Phi^{f,\delta_3}\right)\|_{L^\infty}}{N}.
\end{aligned}$$

By elementary estimates, we therefore conclude that

$$\begin{aligned}
&\left|\int_0^t f(\tau, X(\tau,t,x,\eta),\eta)\,\mathrm{d}\tau - \mathrm{R}\left(\Phi^{f,\mathrm{anti}}\right)(t,x,\eta)\right| \\
&\leq \sup_{(\tau,t,x,\eta)\in[0,T]\times[0,T]\times K\times[0,1]^D)} \left|f^X(\tau,t,x,\eta) - \mathrm{R}\left(\Phi^{f,\delta_3}\odot\Phi^{X,\delta_2,\mathrm{full}}\right)(\tau,t,x,\eta)\right| + \\
&\quad + \frac{2T^2\|f^X\|_{C^1} + 3\|\mathrm{R}\left(\Phi^{f,\delta_3}\right)\|_{L^\infty}}{N} \\
&\underset{(4.5)}{\leq} \mathrm{Lip}_f\delta_2 + \delta_3 + \frac{2T^2\|f^X\|_{C^1} + 3\|\mathrm{R}\left(\Phi^{f,\delta_3}\right)\|_{L^\infty}}{N}.
\end{aligned} \tag{4.7}$$

Setting $\widetilde{X} := X(0,\cdot,\cdot,\cdot)$, we conclude for the NN $\Phi^{u,\varepsilon} := (\Phi^{u_0,\delta_1}\odot\Phi^{\widetilde{X},\delta_2})\oplus\Phi^{f,\mathrm{anti}}$ that

$$\begin{aligned}
&\left\|u - \mathrm{R}\left(\Phi^{u,\varepsilon}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)} \\
&= \left\|u_0\circ\widetilde{X} + \int_0^t f(\tau, X(\tau,t,x,\eta),\eta)\,\mathrm{d}\tau \right.\\
&\quad \left. -\mathrm{R}\left(\Phi^{u_0,\delta_1}\odot\Phi^{\widetilde{X},\delta_2}\right) - \mathrm{R}\left(\Phi^{f,\mathrm{anti}}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)} \\
&\underset{(4.7)}{\leq} \left\|u_0\circ\widetilde{X} - \mathrm{R}(\Phi^{u_0,\delta_1})\circ\mathrm{R}(\Phi^{\widetilde{X},\delta_2})\right\|_{L^\infty([0,T]\times K\times[0,1]^D)} \\
&\qquad +\mathrm{Lip}_f\delta_2 + \delta_3 + \frac{2T^2\|f^X\|_{C^1} + 3\|\mathrm{R}\left(\Phi^{f,\delta_3}\right)\|_{L^\infty}}{N} \\
&\leq \delta_1 + \mathrm{Lip}_{u_0}\delta_2 + \mathrm{Lip}_f\delta_2 + \delta_3 + \frac{2T^2\|f^X\|_{C^1} + 3\|\mathrm{R}\left(\Phi^{f,\delta_3}\right)\|_{L^\infty}}{N} \\
&\leq \frac{\varepsilon}{6} + \frac{\varepsilon}{12} + \frac{\varepsilon}{12} + \frac{\varepsilon}{12} + \frac{\varepsilon}{3} < \varepsilon,
\end{aligned}$$

where we have used (4.6) and (4.4).

We compute the number of weights using Definition 4.1 Theorem 2.6, Propositions 2.2, B.1, and Remark 4.5 as

$$\begin{aligned}
W\left(\Phi^{u,\varepsilon}\right) &\leq W\left(\Phi^{u_0,\delta_1}\odot\Phi^{\widetilde{X},\delta_2}\right) + W\left(\Phi^{f,\mathrm{anti}}\right) \\
&\leq c\cdot(\varepsilon^{-1/r} + \varepsilon^{-d/k})\cdot(\ln(1/\varepsilon)+1) + c'\cdot N\cdot W\left(\Phi^{f,\delta_3}\odot\Phi^{X,\delta_2,\mathrm{full}}\right) \\
&\leq c\cdot\left(\varepsilon^{-1/r} + \varepsilon^{-d/k}\right)\cdot(\ln(1/\varepsilon)+1) + c''\cdot\varepsilon^{-1}\cdot\left(\varepsilon^{-d/k} + \varepsilon^{-(d+1)/k}\right) \\
&\leq c'''\cdot(\varepsilon^{-1/r} + \varepsilon^{-(d+1)/k-1})\cdot(\ln(1/\varepsilon)+1)
\end{aligned}$$

with $c''' = c'''(n,r,d,k,K,T,V,u_0,\|f\|_{C^{s'}}) > 0$. The number of layers is $c''''\cdot(\ln(1/\varepsilon)+1)$, for $c'''' = c''''(n,r,d,k,K,T,V,u_0,\|f\|_{C^{s'}}) > 0$, by the same results. □

4.4 Conservative form

Next, we extend our results to transport equations in conservative form by invoking Proposition 3.10.

Theorem 4.8 *Let V satisfy assumptions (H1) and (H2) with $n, D, k \in \mathbb{N}$, and $T > 0$. Further let, for $r > 0$, $u_0 \in C^1(\mathbb{R}^n)$ be r-approximable by NNs. Let $u \in C^1([0,T] \times \mathbb{R}^n \times [0,1]^D)$ denote the unique solution of the Cauchy problem for the conservative parametric linear transport equation:*

$$\begin{cases} \partial_t u(t,x,\eta) + \operatorname{div}_x(V(t,x,\eta)u(t,x,\eta)) = 0, \\ u(0,x,\eta) = u_0(x). \end{cases}$$

Then, for every $\varepsilon \in (0,1)$ and every compact subset $K \subset \mathbb{R}^n$, there exists a NN $\Phi^{\overline{u},\varepsilon}$ with d-dimensional input, where $d := 1 + n + D$, such that for the restriction $\overline{u} := u\big|_{[0,T]\times K\times[0,1]^D}$ there holds

(i) $L\left(\Phi^{\overline{u},\varepsilon}\right) \leq c \cdot (\ln(1/\varepsilon) + 1)$,
(ii) $W\left(\Phi^{\overline{u},\varepsilon}\right) \leq c \cdot \left(1 + \varepsilon^{-1/r} + \varepsilon^{-d/(k-1)}\right) \cdot (\ln(1/\varepsilon) + 1)$,
(iii) $\left\|\overline{u} - \mathrm{R}\left(\Phi^{\overline{u},\varepsilon}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)} < \varepsilon$,

for $c = c(n, r, d, k, K, T, V, u_0) > 0$.

Remark 4.9
- If $u_0 \in C^s(\mathbb{R}^n)$, then Remark 4.2 demonstrates that we can replace r in Theorem 4.8 by s/n and the constant c in Theorem 4.8 depends more specifically on $\|u_0\|_{C^s}$.
- As in earlier results, the statement of Theorem 4.8 extends to a weak formulation of the transport equation via Remark 3.11.

Proof Recall that, by Proposition 3.10, the unique solution $u \in C^1([0,T] \times \mathbb{R}^n \times [0,1]^D)$ of (3.1) is given by

$$u(t,x,\eta) = u_0(X(0,t,x,\eta))J(0,t,x,\eta).$$

Moreover, $X \in C^k([0,T] \times [0,T] \times \mathbb{R}^n \times [0,1]^D)$ by Theorem 3.3 and therefore $J \in C^{k-1}([0,T] \times [0,T] \times \mathbb{R}^n \times [0,T]^D)$. We define $\widetilde{X} := X(0,.,.,.)$ and $\widetilde{J} := J(0,.,.,.)$.

The proof proceeds by first approximating $\widetilde{J}$ by a NN with help of Theorem 2.6, then invoking the known approximation of $u_0 \circ \widetilde{X}$ via Theorem 4.3, and then applying the multiplication of NNs by Proposition 2.5.

Let $G_1 := \max\{G_0, T\|V\|_{C^1}\exp(T\|V\|_{C^1})\}$ be the bound of $\|\widetilde{X}\|_{C^1}$ from (A.2). The Hadamard inequality [32, Corollary 7.8.2] implies that

$$\begin{aligned} \|\widetilde{J}\|_{L^\infty([0,T]\times K\times[0,1]^D)} &\leq \sup_{(t,x,\eta)\in[0,T]\times K\times[0,1]^D} \prod_{i=1}^{n} \left\|(D_x X(0,t,x,\eta))_i\right\|_2 \\ &\leq \prod_{i=1}^{n} \sqrt{n\|\widetilde{X}\|_{C^1}^2} = n^{n/2}\|\widetilde{X}\|_{C^1}^n \leq n^{n/2}G_1^n =: G_J, \end{aligned}$$

where $(D_x X(s,t,x,\eta))_i$ denotes the i-th column vector of the matrix $D_x X$.
Theorem 2.6 implies that there exist NNs Φ^{u_0,δ_1}, $\Phi^{\widetilde{X},\delta_2}$, and $\Phi^{\widetilde{J},\delta_3}$ with n, d, and d-dimensional input dimension respectively such that, for $\delta_1 := \varepsilon/(8G_J)$, $\delta_2 := \varepsilon/(8\operatorname{Lip}_{u_0} G_J)$, and $\delta_3 := \varepsilon/(4\|u_0\|_{L^\infty})$, there holds

$$\left\|u_0 - \mathrm{R}\left(\Phi^{u_0,\delta_1}\right)\right\|_{L^\infty} < \delta_1, \quad \left\|\widetilde{X} - \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right\|_{L^\infty} < \delta_2, \quad \text{and}$$
$$\left\|\widetilde{J} - \mathrm{R}\left(\Phi^{\widetilde{J},\delta_3}\right)\right\|_{L^\infty} < \delta_3.$$

We conclude for $\Phi^{u,\varepsilon} := (\Phi^{u_0,\delta_1} \odot \Phi^{\widetilde{X},\delta_2}) \otimes^{\varepsilon/4} \Phi^{\widetilde{J},\delta_3}$ that

$$\begin{aligned}
&\left\|u - \mathrm{R}\left(\Phi^{u,\varepsilon}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)}\\
&\quad= \left\|\left(u_0 \circ \widetilde{X}\right)\cdot \widetilde{J} - \mathrm{R}\left(\left(\Phi^{u_0,\delta_1} \odot \Phi^{\widetilde{X},\delta_2}\right) \otimes^{\varepsilon/4} \Phi^{\widetilde{J},\delta_3}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)}\\
&\quad= \left\|\left(u_0 \circ \widetilde{X}\right)\cdot \widetilde{J} - \left(\mathrm{R}\left(\Phi^{u_0,\delta_1}\right) \circ \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right)\cdot \mathrm{R}\left(\Phi^{\widetilde{J},\delta_3}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)} + \frac{\varepsilon}{4}\\
&\quad\le \left\|\left(u_0 \circ \widetilde{X}\right)\cdot \widetilde{J} - \mathrm{R}\left(\Phi^{u_0,\delta_1}\right) \circ \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\cdot \widetilde{J}\right\|_{L^\infty([0,T]\times K\times[0,1]^D)}\\
&\qquad+ \left\|\left(\mathrm{R}\left(\Phi^{u_0,\delta_1}\right) \circ \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right)\cdot \widetilde{J}\right.\\
&\qquad\left.- \left(\mathrm{R}\left(\Phi^{u_0,\delta_1}\right) \circ \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right)\mathrm{R}\left(\Phi^{\widetilde{J},\delta_3}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)} + \frac{\varepsilon}{4}\\
&\quad\le \left\|u_0 \circ \widetilde{X} - \mathrm{R}\left(\Phi^{u_0,\delta_1}\right) \circ \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)} \|\widetilde{J}\|_{L^\infty([0,T]\times K\times[0,1]^D)}\\
&\qquad+ \left\|\mathrm{R}\left(\Phi^{u_0,\delta_1}\right) \circ \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)}\\
&\qquad\times \left\|\widetilde{J} - \mathrm{R}\left(\Phi^{\widetilde{J},\delta_3}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)} + \frac{\varepsilon}{4}\\
&\quad\le \frac{\varepsilon}{4} + \left\|u_0 \circ \widetilde{X} - \mathrm{R}\left(\Phi^{u_0,\delta_1}\right) \circ \mathrm{R}\left(\Phi^{\widetilde{X},\delta_2}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)}\\
&\qquad\times \left\|\widetilde{J} - \mathrm{R}\left(\Phi^{\widetilde{J},\delta_3}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)}\\
&\qquad+ \|u_0 \circ \widetilde{X}\|_{L^\infty([0,T]\times K\times[0,1]^D)} \left\|\widetilde{J} - \mathrm{R}\left(\Phi^{\widetilde{J},\delta_3}\right)\right\|_{L^\infty([0,T]\times K\times[0,1]^D)} + \frac{\varepsilon}{4}\\
&\quad\le \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon.
\end{aligned}$$

We have assumed without loss of generality that $\|\widetilde{J} - \mathrm{R}(\Phi^{\widetilde{J},\delta_3})\|_{L^\infty([0,T]\times K\times[0,1]^D)} \le 1$ and have used the previous result from Theorem (4.3) to bound $\|u_0 \circ \widetilde{X} - \mathrm{R}(\Phi^{u_0,\delta_1}) \circ \mathrm{R}(\Phi^{\widetilde{X},\delta_2})\|_{L^\infty([0,T]\times K\times[0,1]^D)}$.
We compute the number of weights with Definition 4.1, Theorem 2.6, Proposition 2.2, and 2.5 as

$$\begin{aligned}
W\left(\Phi^{u,\varepsilon}\right) &\le c\cdot(\ln(1/\varepsilon)+1) + 2\cdot\left(W\left(\Phi^{u_0,\delta_1} \odot \Phi^{\widetilde{X},\delta_2}\right) + W\left(\Phi^{\widetilde{J},\delta_3}\right)\right)\\
&\le c\cdot\left(1 + \varepsilon^{-1/r} + \varepsilon^{-d/k} + \varepsilon^{d/k-1}\right)\cdot(\ln(1/\varepsilon)+1)
\end{aligned}$$

with $c = c(n,r,d,k,K,T,V,u_0) > 0$. The number of layers is $c\cdot(\ln(1/\varepsilon)+1)$ by the same theorems and propositions. □

5 Extensions

Below, we discuss some natural extensions of our work to more general settings.

- *Bounded domains and boundary conditions:* The previous theory was formulated for $\Omega = \mathbb{R}^n$ to avoid the discussion of boundary conditions and solutions were approximated on compact subsets. Most of the results can be extended to boundary value problems in a straightforward manner. To avoid technicalities in the following discussion, we assume that Ω is a bounded and convex subset of $\mathbb{R}^n$ with smooth boundary. Firstly, we mention the case of a pure characteristic boundary, i.e.:

  $$V(t, x, \eta) \cdot \nu = 0 \quad \forall x \in \partial\Omega, t \in [0, T], \eta \in [0, 1]^D,$$

  where ν denotes the unit outward normal vector of $\partial\Omega$. In this case no boundary conditions have to be prescribed and the solution of (3.1) is given as before by $u_0(X(0, t, x))$; therefore, all the previous results immediately extend to this case. The same holds for periodic boundary conditions on a cube $[0, L]^n$, $L > 0$ with periodic initial conditions since the solution is just the restriction on $[0, L]^n$ of the corresponding Cauchy problem on $\Omega = \mathbb{R}^n$ [1, Remark 2.2.9].
  For inflow boundary conditions, we discuss the case of $V(t, x, \eta) = \tilde{V}(\eta)$ in more detail. The results can be extended to x-dependent vector fields under further geometric assumptions on the vector field which can, for instance, include that any integral curve of the vector field that is tangent to the boundary of the domain remains in the complement of the domain. We define:

  $$\partial\Omega^- := \Big\{(x, \eta) \in \partial\Omega \times [0, 1]^D : \ \tilde{V}(\eta) \cdot \nu < 0\Big\}$$

  and the exit time

  $$\tau_{t,x,\eta} = \inf\big\{s \geq 0 : \ X(s, t, x, \eta) \in \overline{\Omega}\big\}.$$

  Then [1, Theorem 2.2.6] implies that for $u_b^- \in C^1([0, T] \times \partial\Omega^-)$ there holds that the problem

  $$\begin{cases} \partial_t u(t, x, \eta) + \tilde{V}(\eta) \cdot \nabla_x u(t, x, \eta) = 0, \\ u(0, x, \eta) = u_0(x), \\ u(t, x, \eta) = u_b^-(t, x, \eta), \qquad \text{for } (x, \eta) \in \partial\Omega^-, \end{cases} \tag{5.1}$$

  has a unique solution $u \in C^1([0, T] \times \Omega \times [0, 1]^D)$ which is given by

  $$u(t, x, \eta) = \begin{cases} u^0(X(0, t, x, \eta)), & \tau_{t,x,\eta} = 0, \\ u_b^-(\tau_{t,x,\eta}, x^*, \eta), & \tau_{t,x,\eta} > 0, \end{cases} \tag{5.2}$$

  with $x^* = X(\tau_{t,x,\eta}, t, x, \eta)$, if and only if, for all $(y, \eta) \in \partial\Omega^-$ there holds

  $$u_b^-(0, y, \eta) = u_0(y), \quad \partial_t u_b^-(0, y, \eta) + \tilde{V}(\eta) \cdot \nabla u^0(y) = 0.$$

  These results also hold for transport equations that include source terms and amplification factors; see [1, Theorem 2.2.7].

To give an intuition what a NN approximation of (5.2) can look like, we further suppose that u_b^- and u_0 are compatible such that there exists a $\tilde{u} \in C^1([0,T] \times \Omega \times [0,1]^D)$ with

$$\tilde{u}(t,x,\eta) = \begin{cases} u_0(x), & \text{if } t = 0, \\ u_b^-(t,x,\eta), & \text{if } (x,\eta) \in \partial\Omega^-. \end{cases}$$

Then, the solution (5.2) can be written as

$$u(t,x,\eta) = \tilde{u}(\tau_{t,x,\eta}, X(\tau_{t,x,\eta}, t, x, \eta), \eta).$$

Since the solution now has a compositional structure, the concepts we used previously can be used to construct a NN approximation. Note that $\tau_{t,x,\eta}$ is smooth for the prescribed case of a convex domain with smooth boundary and a velocity field independent of x and t. For more general domains, the efficiency of the NN approximation depends mainly on the regularity of the exit time $\tau_{t,x,\eta}$.

- *Non-linear transport equations:* An immediate question is to what extent the results carry over to the non-linear setting. We believe that it is highly unlikely that similar results hold in this regime without overly restrictive assumptions. Indeed, in the non-linear case, non-smoothness of the initial condition u_0 potentially implies non-smoothness of the characteristic curves described by X. This can already be seen in the one-dimensional case. We consider the one-dimensional non-linear transport equation:

$$\begin{cases} \partial_t u(t,x,\eta) + \partial_x[f(u(t,x,\eta))] = 0, \\ u(0,x,\eta) = u_0(x). \end{cases}$$

The characteristic system of ODEs is then given by [53, p. 26] as

$$\left\{ \begin{aligned} &\partial_s X(s,t,x,\eta) = f'(u(t, X(s,t,x,\eta), \eta)), && \text{(5.3a)} \\ &X(t,t,x,\eta) = x. && \text{(5.3b)} \end{aligned} \right.$$

Hence, the regularity of the characteristic curves described by X depends on the global regularity of u and therefore on the regularity of u_0; therefore, X is not guaranteed to be smooth.

If X is non-smooth, then the fundamental backbone of the argument, which is that u can be written as the composition of a high-dimensional smooth and low-dimensional (potentially) rough function, collapses.

- *Non-smooth velocity fields:* As mentioned above, if X cannot be guaranteed to be smooth, our argument cannot be made in the same way as before.

However, it is conceivable that one can have efficient approximations of X by realizations of NNs if X is non-smooth. This is the case if X possesses specific additional structure, such as compositionality. For example, if $X(s,t,x,\eta) = X_1(s,t,x,X_2(\eta))$ with $X_1 : [0,T] \times [0,T] \times \Omega \times [0,1] \to \Omega$, and $X_2 : [0,1]^D \to [0,1]$, where X_2 is smooth but X_1 is non-smooth. Then similar arguments as earlier can be used to establish that X can be efficiently approximated by realizations of NNs.

- *Damping/amplification*: The extension of our results to parametric linear transport equations that include an amplification or damping factor is straightforward. More precisely, we consider solutions of the equation:

$$\begin{cases} \partial_t u(t,x,\eta) + V(t,x,\eta)\cdot\nabla_x u(t,x,\eta) + a(t,x,\eta)u(t,x,\eta) = 0, & (5.4a) \\ u(0,x,\eta) = u_0(x), & (5.4b) \end{cases}$$

  where a is, similarly to (4.1) and (4.2), either given by $a(t,x,\eta) = a_1(t,x) \in C^{s'}([0,T]\times\mathbb{R}^n)$, for $s' := \lceil (n+1)k/d \rceil$, or $a(t,x,\eta) = a_2(t,x,\eta) \in C^{s'}([0,T]\times\mathbb{R}^n\times[0,1]^D)$, for $s' := k$.
  If V satisfies assumptions (H1) and (H2) with $n, D, k \in \mathbb{N}$, $T > 0$, and $u_0 \in C^1(\mathbb{R}^n)$ being r-approximable by NNs for $r > 0$, then one can show, see [25], that there exists a unique solution of (5.4) which is given by

$$u(t,x,\eta) = u_0(X(0,t,x,\eta)) \exp\left(-\int_0^t a(\tau, X(\tau,t,x,\eta),\eta)\,\mathrm{d}\tau\right). \quad (5.5)$$

  To get an estimate on the sizes of approximating NNs for functions of the form of (5.5), we only have to combine previous results. Section 4.3 describes how to approximate the map $(t,x,\eta) \mapsto -\int_0^t a(\tau, X(\tau,t,x,\eta),\eta)\,\mathrm{d}\tau$ by realizations of NNs. This approximation can be concatenated with an approximation of the smooth, one-dimensional exponential function via Proposition 2.2. Finally, the result may be multiplied, via Proposition 2.5, with the already known approximation of $u_0(X(0,t,x,\eta))$ from Theorem 4.3. This yields a NN $\Phi^{\overline{u},\varepsilon}$ such that the realization of $\Phi^{\overline{u},\varepsilon}$ approximates (5.5) up to an error of $\varepsilon > 0$. Estimating the individual sizes of the networks involved in the construction of $\Phi^{\overline{u},\varepsilon}$ yields that

$$\begin{aligned} L\left(\Phi^{\overline{u},\varepsilon}\right) &\le c\cdot(\ln(1/\varepsilon)+1), \\ W\left(\Phi^{\overline{u},\varepsilon}\right) &\le c\cdot\left(\varepsilon^{-1/r} + \varepsilon^{-(d+1)/k-1}\right)\cdot(\ln(1/\varepsilon)+1), \end{aligned}$$

  with $c = c(n,r,d,k,K,T,V,u_0,\|a\|_{C^{s'}}) > 0$. As before, if $u_0 \in C^s(\mathbb{R}^n)$, then $r = s/n$.
- *Parameter dependence of initial condition:* We only considered the case where u_0 does not depend on the parameters. It is not hard to see that, in the framework of r-approximability, the same result would hold if u_0 depended on the parameters. However, if $u_0 \in C^s(\mathbb{R}^n\times[0,1]^D)$, then Remark 4.2 would yield an approximation rate depending on the dimension D of the parameter space.

  For an application of Remark 4.2 it is required that u_0 is a low-dimensional function. Hence, if $u_0 \in C^s$ depends on very few parameters, say the first $t \ll D$, then all main theorems can be extended directly. Instead of approximating $x \mapsto u_0(x)$ with a NN up to an error of $\epsilon > 0$ and having to use $\mathcal{O}(\epsilon^{-n/s})$ many weights for $\varepsilon \to 0$, one would instead approximate $x \mapsto u_0(x,\eta_1,\ldots,\eta_t)$ which requires $\mathcal{O}(\epsilon^{-(n+t)/s})$ many weights for $\varepsilon \to 0$.
  A second framework in which u_0 could be guaranteed to be r-approximable with large r while having low spatial smoothness is that where the parameter dependence is decoupled from the dependence on the spatial coordinates. For example, if $u_0(x,\eta) = \tilde{u}_0(x)\cdot\kappa_1(\eta) + \kappa_2(\eta)$ for smooth κ_1, κ_2, then again low regularity of $\tilde{u}_0$ could suffice to achieve fast rates.

- *Weak solutions with discontinuous initial condition:* Since realizations of deep neural networks are always continuous functions, we cannot hope to obtain approximation results in the uniform norm as studied in this work. However, if one considers L^p-approximation, for $p \in [1, \infty)$ instead, then approximation of piecewise regular functions is possible. This situation was studied in [48].
- *Numerical proof-of-concept:* While we observe theoretically that solutions of certain transport equations can be very well approximated by neural networks with rates virtually independent of the ambient dimension, the asymptotic estimates mask the existence of constants that may be dimension dependent. Hence, it is very well possible that the presented results do not effect practical scenarios. In this context, a comprehensive numerical study should be carried out to analyze the practical effect of the smoothness and the dimensions of parameter spaces and spatial dimension on the approximability of solutions of associated transport equations.

Appendix 1: Bounds for $\|X\|_{C^k}, k = 0, 1$

Proposition A.1 *Let X be defined as in* Theorem 3.3. *Then, for every compact set $K \subset \mathbb{R}^n$ there holds*

$$\|X\|_{C^0([0,T]\times[0,T]\times K\times[0,1]^D)} \leq (|K| + CT)\exp(CT) := G_0 \tag{A.1}$$

$$\|X\|_{C^1([0,T]\times[0,T]\times K\times[0,1]^D)} \leq \max\left\{G_0, T\|V\|_{C^1}\exp(T\|V\|_{C^1})\right\} \tag{A.2}$$

with $\|V\|_{C^1} := \|V\|_{C^1([0,T]\times B_{G_0}(0)\times[0,1]^D)}$.

Proof We start with the definition of X given by

$$\begin{cases} \partial_s X(s,t,x,\eta) = V(s, X(s,t,x,\eta), \eta), & \text{(A.3a)} \\ X(t,t,x,\eta) = x. & \text{(A.3b)} \end{cases}$$

The fundamental theorem of calculus implies

$$X(s,t,x,\eta) = x + \int_t^s V(\tau, X(\tau,t,x,\eta), \eta)\,\mathrm{d}\tau. \tag{A.4}$$

With the help of the sub-linear growth-condition (H2), we conclude

$$\begin{aligned} |X(s,t,x,\eta)| &\leq |x| + \left|\int_t^s |V(\tau, X(\tau,t,x,\eta),\eta)|\,\mathrm{d}\tau\right| \\ &\leq |x| + C\left|\int_t^s (1 + |X(\tau,t,x,\eta)|)\,\mathrm{d}\tau\right| \\ &\leq |x| + CT + C\int_t^s |X(\tau,t,x,\eta)|\,\mathrm{d}\tau. \end{aligned}$$

Moreover, by Gronwall's inequality

$$\sup_{s\in[0,T]} |X(s,t,x,\eta)| \leq (|x| + CT)\exp(CT).$$

Hence,

$$\|X\|_{C^0([0,T]\times[0,T]\times K\times[0,1]^D)} \le (|K| + CT)\exp(CT).$$

We have by (A.3a)

$$\|\partial_s X\|_{C^0} \le \|V\|_{C^0}.$$

Furthermore, applying Leibniz integral rule to (A.4) yields

$$\partial_t X(s,t,x,\eta) = -V(t,x,\eta) + \int_t^s \nabla_x V(\tau, X(\tau,t,x,\eta),\eta)\partial_t X(\tau,t,x,\eta)\,\mathrm{d}\tau$$

and therefore

$$|\partial_t X(s,t,x,\eta)| \le \|V\|_{C^0} + \|V\|_{C^1}\int_t^s |\partial_t X(\tau,t,x,\eta)|\,\mathrm{d}\tau.$$

Gronwall's inequality implies then

$$\|\partial_t X\|_{C^0} \le \|V\|_{C^0}\exp(T\|V\|_{C^1}).$$

The same procedure results for $\nabla_x X$ and $\nabla_\eta X$ in

$$\begin{aligned}\|\nabla_x X\|_{C^0} &\le \exp(T\|V\|_{C^1}),\\ \|\nabla_\eta X\|_{C^0} &\le T\|V\|_{C^1}\exp(T\|V\|_{C^1}).\end{aligned}$$

Thus, we get after assuming without loss of generality that $T \ge 1$, $\|V\|_{C^1} \ge 1$

$$\|X\|_{C^1} \le \max\left\{G_0, T\|V\|_{C^1}\exp(T\|V\|_{C^1})\right\}. \tag{A.5}$$

□

Appendix 2: Construction of a NN emulating the left Riemann sum

Proposition B.1 *Let $d \in \mathbb{N}_{\ge 2}$, $T > 0$, $\Omega \subset \mathbb{R}^{d-1}$, and let Φ be a NN with d-dimensional input. Then there exists a NN $\widetilde{I}_N(\Phi)$ such that*

- $L\left(\widetilde{I}_N(\Phi)\right) = L(\Phi) + c_1$,
- $W\left(\widetilde{I}_N(\Phi)\right) \le c_2 \cdot N \cdot W(\Phi)$,
- $$\sup_{t\in[0,T],x\in\Omega}\left|\mathrm{R}\left(\widetilde{I}_N(\Phi)(t,x)\right) - \frac{1}{N}\sum_{i=0}^{\lceil tN/T\rceil-1}\mathrm{R}(\Phi)\left(\frac{iT}{N},x\right)\right| \le \frac{c_3}{N}, \tag{B.1}$$

where $c_1, c_2 > 0$ are independent of Φ and $c_3 := 3\|\mathrm{R}(\Phi)\|_{L^\infty([0,T]\times\Omega)}$.

Proof Let, for $i \in \{0,1,\dots,N\}$, $t_i := iT/N$. We define, for $i \in \{0,\dots,N-1\}$,

$$\Phi_i^{(\mathrm{shift})} := \Phi \odot \left(\begin{pmatrix}0 & 0\\ 0 & \mathrm{Id}_{\mathbb{R}^{d-1}}\end{pmatrix}, \begin{pmatrix}t_i\\ 0\end{pmatrix}\right).$$

Then $\mathrm{R}(\Phi_i^{(\mathrm{shift})})(t,x) = \mathrm{R}(\Phi)(t_i,x)$, for all $t \in [0,T]$, $x \in \Omega$. Moreover, $W(\Phi_i^{(\mathrm{shift})}) \leq 2W(\Phi) + 2d$ and $L(\Phi_i^{(\mathrm{shift})}) = L(\Phi) + 2$ by Proposition 2.2. Next, we define the following indicator networks for $i \in \{0,\ldots,N-1\}$:

$$\Phi_i^{(\mathrm{ind})} := \left(([1 \quad 0 \quad \ldots \quad 0], 0), \left(\left[\begin{array}{c|c} 1 & 0_{\mathbb{R}^d} \\ 1 & 0_{\mathbb{R}^d} \end{array} \right], \begin{pmatrix} -t_i \\ -t_{i+1} \end{pmatrix} \right), \left(\left[N \quad -N \quad | \quad 0 \quad | \quad 0_{\mathbb{R}^d} \right], 0 \right) \right).$$

We have that $W(\Phi_i^{(\mathrm{ind})}) = 7$, $L(\Phi_i^{(\mathrm{ind})}) = 3$ and, for $t \in [0,T]$ and $x \in \Omega$,

$$\mathrm{R}\left(\Phi_i^{(\mathrm{ind})}\right)(t,x) = N \cdot (\varrho(t-t_i) - \varrho(t-t_{i+1})) = \begin{cases} 0 & \text{if } t \leq t_i, \\ N \cdot (t-t_i) & \text{if } t_i < t < t_{i+1}, \\ 1 & \text{if } t \geq t_{i+1}. \end{cases} \tag{B.2}$$

Let $\bar{a} := \|\mathrm{R}(\Phi)\|_{L^\infty([0,T]\times\Omega)}$. Now we set, for $i \in \{0,\ldots,N-1\}$,

$$\Phi_i^{(\mathrm{clip})} := \left(\left(\begin{pmatrix} 2\bar{a} & 1 \\ 2\bar{a} & 0 \end{pmatrix}, \begin{pmatrix} -\bar{a} \\ -\bar{a} \end{pmatrix} \right), ([1 \quad -1 \quad | \quad 0 \quad 0], 0) \right) \odot \mathrm{P}\left(\Phi_i^{(\mathrm{ind})}, \Phi_i^{(\mathrm{shift})}\right).$$

We have that

$$\begin{aligned} \mathrm{R}\left(\Phi_i^{(\mathrm{clip})}\right)(t,x) &= \varrho\left(2\bar{a}\mathrm{R}\left(\Phi_i^{(\mathrm{ind})}\right)(t,x) + \mathrm{R}\left(\Phi_i^{(\mathrm{shift})}\right)(t,x) - \bar{a}\right) \\ &\quad - \varrho\left(2\bar{a}\mathrm{R}\left(\Phi_i^{(\mathrm{ind})}\right)(t,x) - \bar{a}\right). \end{aligned} \tag{B.3}$$

It follows from (B.2) and (B.3) that, for $t \in [0,T]$ and $x \in \Omega$,

$$\mathrm{R}\left(\Phi_i^{(\mathrm{clip})}\right)(t,x) = 0, \ \text{if } t \leq t_i \tag{B.4}$$

$$\mathrm{R}\left(\Phi_i^{(\mathrm{clip})}\right)(t,x) = \mathrm{R}\left(\Phi_i^{(\mathrm{shift})}\right)(t,x), \ \text{if } t \geq t_{i+1}, \tag{B.5}$$

$$\left|\mathrm{R}\left(\Phi_i^{(\mathrm{clip})}\right)(t,x)\right| \leq 2\bar{a}, \ \text{else}. \tag{B.6}$$

In addition, by Propositions 2.2 and 2.3,

$$L\left(\Phi_i^{(\mathrm{clip})}\right) = 2 + \max\{3, L(\Phi) + 2\}, \tag{B.7}$$

$$W\left(\Phi_i^{(\mathrm{clip})}\right) \leq 16 + 2 \cdot (7 + 2W(\Phi) + 2d). \tag{B.8}$$

Finally, we set

$$\widetilde{I}_N(\Phi) := \left(\left(\left[\frac{1}{N} \quad \ldots \quad \frac{1}{N} \right], 0 \right) \right) \odot \mathrm{P}\left(\Phi_0^{(\mathrm{clip})}, \ldots, \Phi_{N-1}^{(\mathrm{clip})}\right).$$

Now we have that $t \leq t_i$ if $\lceil tN/T \rceil \leq i$ and $t \geq t_{i+1}$ if $i \leq \lceil tN/T \rceil - 1$. Hence, for $t \in [0, T]$ and $x \in \Omega$,

$$\begin{aligned}
\mathrm{R}\left(\widetilde{I}_N(\Phi)\right)(t,x) &= \frac{1}{N}\sum_{i=0}^{N-1} \mathrm{R}\left(\Phi_i^{(\mathrm{clip})}\right)(t,x) \\
&\underset{(\mathrm{B.4})}{=} \frac{1}{N}\sum_{i=0}^{\lceil tN/T \rceil -1} \mathrm{R}\left(\Phi_i^{(\mathrm{clip})}\right)(t,x) \\
&\underset{(\mathrm{B.5})}{=} \frac{1}{N}\sum_{i=0}^{\lceil tN/T \rceil -2} \mathrm{R}\left(\Phi_i^{(\mathrm{shift})}\right)(t,x) + \frac{1}{N}\mathrm{R}\left(\Phi_{\lceil tN/T \rceil -1}^{(\mathrm{clip})}\right)(t,x) \\
&= \frac{1}{N}\sum_{i=0}^{\lfloor tN/T \rfloor -2} \mathrm{R}(\Phi)(t_i,x) + \frac{1}{N}\mathrm{R}\left(\Phi_{\lceil tN/T \rceil -1}^{(\mathrm{clip})}\right)(t,x) \\
&= \frac{1}{N}\sum_{i=0}^{\lceil tN/T \rceil -1} \mathrm{R}(\Phi)(t_i,x) \\
&\quad + \frac{1}{N}\left(\mathrm{R}\left(\Phi_{\lceil tN/T \rceil -1}^{(\mathrm{clip})}\right)(t,x) - \mathrm{R}(\Phi)\left(t_{\lceil tN/T \rceil -1},x\right)\right).
\end{aligned}$$

Since, by (B.6),

$$\frac{1}{N}\left|\mathrm{R}\left(\Phi_{\lceil tN/T \rceil -1}^{(\mathrm{clip})}\right)(t,x) - \mathrm{R}(\Phi)\left(t_{\lceil tN/T \rceil -1},x\right)\right| \leq 3\|\mathrm{R}(\Phi)\|_{L^\infty([0,T]\times\Omega)},$$

we conclude the proof by observing with (B.7) and (B.8) that

$$\begin{aligned}
L\left(\widetilde{I}_N(\Phi)\right) &\leq 3 + \max\{3, L(\Phi) + 2\}, \\
W\left(\widetilde{I}_N(\Phi)\right) &\leq 2N + N \cdot (32 + 4 \cdot (7 + 2W(\Phi) + 2d)) \\
&= 62N + 8W(\Phi)N + 8dN.
\end{aligned}$$

□

Proposition B.2 (Left Riemann sum) *Let $M > 0$, $n \in \mathbb{N}$, $f \in C^1([0,T] \times [-M,M]^n; \mathbb{R})$, and $N \in \mathbb{N}$. The approximation of the integral of f with respect to its first argument from 0 to $t \leq T$, $T \geq 1$ by the left Riemann sum is given by*

$$I_N(f)(t,x) := \frac{1}{N}\sum_{i=0}^{N-1} f(t_i,x)\mathbb{1}_{\{t_i<t\}}, \qquad t_i = iT/N.$$

Then

$$\sup_{t\in[0,T],x\in[-M,M]^n}\left|\int_0^t f(\tau,x)\,\mathrm{d}\tau - I_N(f)(t,x)\right| \leq \frac{2T^2}{N}\|f\|_{C^1}.$$

Proof Let $N(t) := \max\{i \in \mathbb{N} \mid t_i < t\}$. Then

$$\begin{aligned}
&\left|\int_0^t f(\tau, x)\,\mathrm{d}\tau - I_N(f)(t, x)\right| \\
&= \left|\sum_{i=0}^{N(t)} \int_{t_i}^{t_{i+1}} f(\tau, x) - f(t_i, x)\,\mathrm{d}\tau - \int_t^{t_{N(t)+1}} f(\tau, x)\,\mathrm{d}\tau\right| \\
&\le \left|\sum_{i=0}^{N(t)} \int_{t_i}^{t_{i+1}} f(\tau, x) - f(t_i, x)\,\mathrm{d}\tau\right| + \left|\int_t^{t_{N(t)+1}} f(\tau, x)\,\mathrm{d}\tau\right| \\
&\le \frac{T^2}{N}\|f\|_{C^1} + \frac{T}{N}\|f\|_{C^0} \le \frac{2T^2}{N}\|f\|_{C^1}.
\end{aligned}$$

□

Remark B.3 Equation (B.1) implies that for a NN Φ with $n+1$-dimensional input there holds

$$\sup_{t\in[0,T], x\in[-M,M]^n} \left|\mathrm{R}\left(\widetilde{I}_N(\Phi)(t,x)\right) - I_N(\mathrm{R}(\Phi))(t,x)\right| \le \frac{c}{N}$$

with $c = 3\|\mathrm{R}(\Phi)\|_{L^\infty([0,T]\times\Omega)} > 0$.

Acknowledgments The authors would like to thank Avi Mayorcas for inspiring discussions in the early stage of this work. The authors thank Francois Golse for advice and various helpful suggestions on the theory of transport equations. P.P is grateful for the hospitality and support of the Institute of Mathematics of the University of Oxford during his visit in January 2020.

Funding Open Access funding provided by University of Vienna.

References

1. Allaire, G., Blanc, X., Després, B., Golse, F.: Transport et diffusion. Ecole Polytechnique (2019)
2. Ambrosio, L.: Transport equation and cauchy problem for non-smooth vector fields. In: Calculus of Variations and Nonlinear Partial Differential Equations, pp. 1–41. Springer, Berlin (2008)
3. Ambrosio, L., Gigli, N., Savare, G.: Gradient Flows. Birkhäuser-Verlag (2005)
4. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory **39**(3), 930–945 (1993)
5. Beck, C., Becker, S., Grohs, P., Jaafari, N., Jentzen, A.: Solving stochastic differential equations and kolmogorov equations by means of deep learning. arXiv:1806.00421 (2018)
6. Beck, C., Jentzen, A., Kuckuck, B.: Full error analysis for the training of deep neural networks. arXiv:1910.00121 (2019)

7. Bellman, R.: On the theory of dynamic programming. Proc. Natl. Acad. Sci. U.S.A. **38**(8), 716 (1952)
8. Berner, J., Grohs, P., Jentzen, A.: 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:1809.03062 (2018)
9. Bölcskei, H., Grohs, P., Kutyniok, G., Petersen, P.: Optimal approximation with sparsely connected deep neural networks. SIAM Journal on Mathematics of Data Science **1**(1), 8–45 (2019)
10. Bouchut, F., Golse, F., Pulvirenti, M.: Kinetic Equations and Asymptotic Theory. Series in Applied Mathematics, 4. Gauthier-Villars, Editions Scientifiques et M'edicales Elsevie (2000)
11. Chen, M., Jiang, H., Liao, W., Zhao, T.: Efficient approximation of deep relu networks for functions on low dimensional manifolds. In: Advances in Neural Information Processing Systems, pp. 8172–8182 (2019)
12. Courant, R., Hilbert, D.: Methods of Mathematical Physics. Wiley, New York (1989)
13. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems **2**(4), 303–314 (1989)
14. Dahmen, W., Gruber, F., Mula, O.: An adaptive nested source term iteration for radiative transfer equations. Mathematics of Computation (2020)
15. Dahmen, W., Huang, C., Kutyniok, G., Lim, W.-Q., Schwab, C., Welper, G.: Efficient resolution of anisotropic structures. In: Extraction of Quantifiable Information from Complex Systems, pp. 25–51. Springer (2014)
16. Dahmen, W., Kutyniok, G., Lim, W.-Q., Schwab, C., Welper, G.: Adaptive anisotropic Petrov–Galerkin methods for first order transport equations. J. Comput. Appl. Math. **340**, 191–220 (2018)
17. Dahmen, W., Plesken, C., Welper, G.: Double greedy algorithms: reduced basis methods for transport dominated problems. ESAIM: Mathematical Modelling and Numerical Analysis **48**(3), 623–663 (2014)
18. DiPerna, R.J., Lions, P.L.: Ordinary differential equations, transport theory and Sobolev spaces. Invent. Math. **98**(3), 511–547 (1989)
19. E, W., Yu, B.: The deep ritz method: a deep learning-based numerical algorithm for solving variational problems. Communications in Mathematics and Statistics **6**(1), 1–12 (2018)
20. Egger, H., Schlottbom, M.: A mixed variational framework for the radiative transfer equation. Mathematical Models and Methods in Applied Sciences **22**(03), 1150014 (2012)
21. Elbrächter, D., Grohs, P., Jentzen, A., Schwab, C.: DNN expression rate analysis of high-dimensional PDEs: application to option pricing. arXiv:1809.07669 (2018)
22. Evans, L.: Partial differential equations. American Mathematical Society (2010)
23. Fresca, S., Dede, L., Manzoni, A.: A comprehensive deep learning-based approach to reduced order modeling of nonlinear time-dependent parametrized PDEs. arXiv:2001.04001 (2020)
24. F.Golse.: Distributions, analyse de Fourier, équations aux dérivées partielles. Ecole polytechnique (2012)
25. Golse, F.: Lecture notes on mean field kinetic equations. https://www.cmls.polytechnique.fr/perso/golse/M2/PolyKinetic.pdf (2013)
26. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016). http://www.deeplearningbook.org
27. Grella, K.: Sparse tensor approximation for radiative transport. PhD thesis, ETH Zurich (2013)
28. Gühring, I., Kutyniok, G., Petersen, P.: Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. Analysis and Applications (in Press)
29. Hartman, P.: Ordinary differential equations. Society for Industrial and Applied Mathematics (2002)
30. He, J., Li, L., Xu, J., Zheng, C.: ReLU deep Neural Networks and Linear Finite Elements. arXiv:1807.03973 (2018)
31. Hesthaven, J.S., Rozza, G., Stamm, B., et al.: Certified Reduced Basis Methods for Parametrized Partial Differential Equations, vol. 590. Springer, Berlin (2016)
32. Horn, R.A., Johnson, C.R. (eds.): Matrix Analysis. Cambridge University Press, Cambridge (1985)
33. Hornik, K., Stinchcombe, M., White, H., et al.: Multilayer feedforward networks are universal approximators. Neural Networks **2**(5), 359–366 (1989)
34. Hutzenthaler, M., Jentzen, A., Kruse, T., Nguyen, T.A.: A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. arXiv:1901.10854 (2019)
35. John, F.: Partial differential equations. Springer, US (1978)

36. Kutyniok, G., Petersen, P., Raslan, M., Schneider, R.: A theoretical analysis of deep neural networks and parametric PDEs. arXiv:1904.00377 (2019)
37. Laakmann, F.: A theoretical analysis of high-dimensional parametric transport equations and neural networks. Project thesis, University of Oxford (2019)
38. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
39. Mhaskar, H.N.: Approximation properties of a multilayered feedforward artificial neural network. Adv. Comput. Math. **1**(1), 61–80 (1993)
40. Mhaskar, H.N.: Neural networks for optimal approximation of smooth and analytic functions. Neural Comput. **8**(1), 164–177 (1996)
41. Montanelli, H., Yang, H., Du, Q.: Deep ReLU networks overcome the curse of dimensionality for bandlimited functions. arXiv:1903.00735 (2019)
42. Mouhot, C.: Hyperbolicity: scalar transport equations, wave equations. University of Cambridge. https://cmouhot.files.wordpress.com/1900/10/chapter41.pdf (2013)
43. Novak, E., Woźniakowski, H.: Approximation of infinitely differentiable multivariate functions is intractable. J. Complex. **25**(4), 398–404 (2009)
44. Obermeier, A., Grohs, P.: On the approximation of functions with line singularities by ridgelets. Journal of Approximation Theory **237**, 30–95 (2019)
45. Ohlberger, M., Rave, S.: Reduced basis methods: success, limitations and future challenges. arXiv:1511.02021 (2015)
46. Opschoor, J.A., Petersen, P., Schwab, C.: Deep ReLU networks and high-order finite element methods. Analysis and Applications (in Press)
47. Opschoor, J.A.A., Schwab, C., Zech, J.: Exponential ReLU DNN expression of holomorphic maps in high dimension. Technical Report 2019-35, Seminar for Applied Mathematics, ETH Zürich, Switzerland (2019)
48. Petersen, P., Voigtlaender, F.: Optimal approximation of piecewise smooth functions using deep reLU neural networks. Neural Netw. **108**, 296–330 (2018)
49. Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., Liao, Q.: Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. Int. J. Autom. Comput. **14**(5), 503–519 (2017)
50. Quarteroni, A., Rozza, G., et al.: Reduced Order Methods for Modeling and Computational Reduction, vol. 9. Springer, Berlin (2014)
51. Schmidt-Hieber, J.: Deep ReLU network approximation of functions on a manifold. arXiv:1908.00695 (2019)
52. Schwab, C., Zech, J.: Deep learning in high dimension: neural network expression rates for generalized polynomial chaos expansions in UQ. Anal. Appl. **17**(01), 19–55 (2019)
53. Serre, D.: Systems of Conservation Laws 1. Cambridge University Press (1999)
54. Shaham, U., Cloninger, A., Coifman, R.R.: Provable approximation properties for deep neural networks. Appl. Comput Harmon. Anal. **44**(3), 537–557 (2018)
55. Sirignano, J., Spiliopoulos, K.: DGM: a deep learning algorithm for solving partial differential equations. J. Comput. Phys. **375**, 1339–1364 (2018)
56. Suzuki, T.: Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. arXiv:1810.08033 (2018)
57. Yarotsky, D.: Error bounds for approximations with deep reLU networks. Neural Netw. **94**, 103–114 (2017)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.