Constructive Approximation
https://doi.org/10.1007/s00365-024-09699-9

CONSTRUCTIVE
APPROXIMATION

Compositional Sparsity, Approximation Classes, and Parametric Transport Equations

Wolfgang Dahmen[1]

Received: 15 July 2022 / Revised: 1 May 2024 / Accepted: 23 October 2025

Abstract
Approximating functions of a large number of variables poses particular challenges often subsumed under the term "Curse of Dimensionality" (CoD). Unless the approximated function exhibits a very high level of smoothness the CoD can be avoided only by exploiting some typically hidden *structural sparsity*. In this paper we propose a general framework for new model classes of functions in high dimensions. They are based on suitable notions of *compositional dimension-sparsity* quantifying, on a continuous level, approximability by compositions with certain structural properties. In particular, this describes scenarios where deep neural networks can avoid the CoD. The relevance of these concepts is demonstrated for *solution manifolds* of parametric transport equations. For such PDEs parameter-to-solution maps do not enjoy the type of high order regularity that helps to avoid the CoD by more conventional methods in other model scenarios. Compositional sparsity is shown to serve as the key mechanism for proving that sparsity of problem data is inherited in a quantifiable way by the solution manifold. In particular, one obtains convergence rates for deep neural network realizations showing that the CoD is indeed avoided.

Keywords Tamed compositions · Compositional approximation · Approximation classes · Deep neural networks · Operator learning · Parametric transport equations · Solution manifolds · Nonlinear widths

Mathematics Subject Classification 41A25 · 35A35 · 41A63 · 35B30 · 35L04 · 41A46

Dedicated to Ronald DeVore on the occasion of his 80th birthday.

Communicated by Zuowei Shen.

This work has been supported in part by National Science Foundation grants NSF-DMS-2012469, NSF-DMS-2038080, NSF-DMS-2245097, and the SFB 1481, funded by the German Research Foundation.

✉ Wolfgang Dahmen
DAHMEN@math.sc.edu

[1] Mathematics Department, University of South Carolina, Columbia, SC, USA

Published online: 25 January 2025

1 What this Is About

We first aim to destill several questions that guide the subsequent discussion in this paper.

1.1 High Dimensional Approximation, Solution Manifolds, and Operator Learning

The need to approximate functions of a large number of variables arises in variety of contexts, most notably maschine learning, partial differential equations (PDEs) like Schrödinger equations, Kolmogorov equations, Fokker-Planck equations, describing the evolution of probability distributions in high dimensional phase spaces, see e.g [3]. Third, and this will be a focus in this paper, in *Uncertainty Quantification* (UQ) one typically deals with PDEs (in low dimension) whose coefficients, constituitive laws or other input-data depend on parameters whose range is to ensure that corresponding parameter dependent solutions cover the range of physically viables states of interest. In abstract terms this can be formulated as: for each parameter (vector) $y \in \mathcal{Y}$, a typically high dimensional parameter domain, find a solution u in a suitable trial space $\mathbb{X}$ such that

$$\mathcal{F}(u, y) = 0. \tag{1.1}$$

Here $\mathcal{F}(\cdot, y)$ stands for a Differential operator in residual form that depends on $y \in \mathcal{Y}$. Thus, a solution $u = u(x, y)$ becomes a function of the spatio-temporal variables x (or (t, x) if one wishes to distinguish the time variable from the spatial ones) in a computational domain Ω, say, as well as on the parametric variable $y \in \mathcal{Y}$. Applications concerning this setting require exploring the corresponding *solution manifold*

$$\mathcal{M} := \{u = u(y) \in \mathbb{X} : y \in \mathcal{Y}\},$$

comprised of all states in $\mathbb{X}$ that arise when traversing the parameter domain $\mathcal{Y}$. In essence this amounts to efficiently evaluating the *parameter-to-solution* map $y \mapsto u(y) \in \mathbb{X}$, for instance, by generating *surrogates* for this map. Hence, this boils down to approximating functions over the high dimensional domain $\Omega \times \mathcal{Y}$. Determining such surrogates can be interpreted as *nonlinear model reduction* or *operator learning*, depending on how the approximation is fabricated. The latter viewpoint is a currently vibrant research field, see [21] and the references cited there.

We adopt here a wider notion of solution manifold that arises in the context of operator learning. Rather than confining the discussion to finite dimensional parameter domains $\mathcal{Y} \subset \mathbb{R}^{d_y}$, $\mathcal{Y}$ could be viewed as a class of functions, accommodating, for instance, initial or boundary conditions, or right hand sides, which can be viewed as *infinite paramater arrays*. In computations the elements of $\mathcal{Y}$ need to be approximated as well, (or truncated) in balance with the overall target accuracy. In the present paper we keep $\mathcal{Y}$ finite dimensional and use a different notation for those problem data that are infinite dimensional fields (see Theorem 4.9). At any rate, *operator learning* is therefore intrinsically high (even infinite) dimensional.

A notorious obstruction encountered in any high-dimensional approximation task is the so-called *Curse of Dimensionality* (CoD). It roughly expresses an exponential

dependence of computational complexity on the number of variables the approximand depends on. One should keep in mind, that one can quantify approximation performance in $\mathbb{X}$ only for elements of some *compact model class* $\mathcal{K}$ in $\mathbb{X}$. Thus, being able to avoid the CoD or not, depends on *both*, the *model class* of functions one wishes to approximate in quantifiable terms, and the *approximation system* (polynomials, rational functions, splines, wavelet systems, frames and even more general dictionaries, or neural networks) one wishes to employ for this purpose. This raises the question:

(Q): which model classes capture the properties of solution manifolds well and what are suitable approximation systems?

This issue is relevant for describing both the domain and the range of (solution) operators in operator learning.

A powerful concept for understanding the interplay between model class and approximation systems is the notion of *approximation classes*. Roughly, $\mathcal{A}^s = \mathcal{A}^s(\Sigma_n, \mathbb{X})$ collects all elements in a Banach space $\mathbb{X}$ that can be approximated in $\mathbb{X}$ by elements from the given approximation system Σ_n with a rate $O(n^{-s})$. A central theme in approximation theory is then to characterize $\mathcal{A}^s$ by *intrinsic properties* of its elements like Sobolev- or Besov-regularity. This has been very successful for systems based on *spatial localization* (splines, wavelets) (nearly) characterizing the classes $\mathcal{A}^s$ as such smoothness spaces $\mathbb{X}^s$ suggesting corresponding unit balls $\mathcal{K} = U\mathbb{X}^s$ as suitable model classes, see [9]. A prototype estimate for the best approximation of a function $v \in \mathbb{X}^s$, a subspace $\mathbb{X}^s$ of smoothness s in $\mathbb{X}$, by piecewise polynomials of order at least s on partitions with N cells reads

$$\|v - v_N\|_{\mathbb{X}} \leq CN^{-s/d}|v|_{\mathbb{X}^s}, \quad N \in \mathbb{N}. \tag{1.2}$$

Thus, to achieve target accuracy $\varepsilon > 0$ requires the order of $\varepsilon^{-d/s}$ degrees of freedom, which is intractable for large d and fixed s. In brief, using Sobolev- or Besov-balls in high dimensions as model classes in conjunction with localization based approximation systems is subject to the CoD. The diminished relevance of classical smoothness in high dimensions is further underlined by the results in [28] saying that the necessary number of functional evaluations of a function f in $C^\infty([0,1]^d)$, with all derivatives bounded by one, to recover f within a given tolerance, is exponential in d. As a consequence, regardless of the type of the operator, operator learning is a priori doomed to suffer from the CoD if the model classes for the data fields have in essence the same entropy numbers as classical smoothness classes in high dimensions. In that sense one could say that a principal shortcoming of the current theory on operator learning is the lack of appropriate model classes for the operator domain that, on the one hand are "thin" enough to avoid the CoD, but still fit relevant application scenarios—one of the central objectives in this paper.

Thus, (Q) requires looking for different types of model classes. One angle is to determine the *Kolmogorov n-widths* of $\mathcal{M}$. More generally, for a given compact set $\mathcal{K} \subset \mathbb{X}$ (the model class) in a Banach space $\mathbb{X}$ they are defined as

$$d_n(\mathcal{K})_{\mathbb{X}} := \inf_{\mathrm{V}_n = n} \sup_{v \in \mathcal{K}} \inf_{v_n \in V_n} \|v - v_n\|_{\mathbb{X}},$$

quantifying how well each $v \in \mathcal{K}$ can be approximated by elements from a *single linear* space.

Addressing a central question in Uncertainty Quantification (UQ), it turns out that when (1.1) is a second order elliptic model where the diffusion coefficients depend *affinely* on the parameters, $d_n(\mathcal{M})_{H^1(\Omega)}$ decays exponentially and, under suitable summability conditions of such parameter expansions, even *robustly* in the parameter dimension d_y. In fact, it can be shown that under such circumstances the map $y \mapsto u(y)$ is even *holomorphic*, see [8] and the references cited there. Thus, at least indirectly very high smoothness comes to aid and approximation systems like sparse high dimensional polynomial expansions in y or Reduced Basis methods can be shown to avoid the CoD, see e.g. [4].

Unfortunately, for a wide range of PDE models (1.1) holomorphy of $y \mapsto u(y)$ does not persist to hold and the Kolmogorov widths exhibit a poor dimension dependent decay. This is, in particular, the case for non-dissipative models describing wave propagation or transport, see Sect. 1.3 for more details.

Apparently, for such scenarios *linear approximation*, i.e., approximating the objects in a model class from a single linear space (although determining such a space and a concrete approximation from that space could be a nonlinear process) is no longer adequate. Instead, *nonlinear approximation*, i.e., the approximation system consists of functions that depend in a nonlinear way on its degrees of freedom, as a basis for *nonlinear model reduction*, might offer advantages. A prominent example of a highly nonlinear approximation systems are (Deep) Neural Networks ((D)NNs) which play a central role also in what follows.

To put this into proper perspective, as a benchmark for the ability of approximating a compact set (a model class) by elements from a nonlinear system, several variants of *nonlinear* n-widths have been proposed, based on factoring the construction of a nonlinear approximation as follows

$$\delta_n(\mathcal{K})_{\mathbb{X}} = \inf_{E_n, D_n} \sup_{v \in \mathcal{K}} \|v - D_n(E_n v)\|_{\mathbb{X}}. \tag{1.3}$$

Here the infimum is taken over all encoder-decoder pairs $E_n : \mathcal{K} \to \mathbb{R}^n$, $D_n : \mathbb{R}^n \to \mathbb{X}$. To exclude practically meaningless pairs, for instance, based on space filling curves, one subjects E_n, D_n to stability conditions whose choice gives rise to several variants of such widths, such as manifold widths, stable withds, Lipschitz widths, see [5, 11, 30]. The stability requirements in the latter two variants are somewhat stronger than mere continuity originally required for manifold widths in [11]. As a consequence, one is able to establish the validity of a version of Carl's inequality. This means that a certain decay rate of those nonlinear widths implies a certain decay rate of the *entropy numbers* $e_n(\mathcal{K})_{\mathbb{X}}$. They mark the smallest radius that can be attained by covers of $\mathcal{K}$ consisting of at most 2^n $\mathbb{X}$-balls (which means that the centers of such covers can be encoded by n-bits), see e.g. [9, 26].

A first interesting consequence of these findings is that, if one insists on one of the above stability requirements on nonlinear approximation the possible gain of nonlinear approximation over linear approximation is in some sense sandwiched by the discrepancy between the respective rates of entropy numbers and Kolmogorov widths.

In particular, when the entropy numbers already suffer from the CoD, as is the case for the classical smoothness based model classes $U\mathbb{X}^s$, nonlinear approximation offers little benefit. This motivates the following refinement of question (Q):

(Q*): What are model classes $\mathcal{K}$ in high dimensions d, that are in the above sense relevant for approximating solution manifolds, and for which nonlinear approximation can avoid the CoD while linear approximation can't. Specifically, for which model class there holds for instance

$$d_n(\mathcal{K})_{\mathbb{X}} = O(n^{-s/d}) \quad \text{while} \quad \delta_n(\mathcal{K})_{\mathbb{X}} = O(d^{\alpha} n^{-\beta}), \quad n \to \infty,$$

where β is independent of d? Specifically, what is the role of DNNs in this context?

1.2 Approximation by DNNs and New Model Classes

There is by now a substantial body of work on the approximation properties of shallow and deep neural networks, in modern jargon on their *expressive power*. A wide range of results is obtained by "emulation" using that the building blocks employed in conventional systems permit (exponentially accurate) representations in terms of small neural networks. These findings show in essence that (perhaps up to log-factors) DNN approximation can match the performance of a diversity of conventional methods, meeting in particular the n-widths benchmark in numerous scenarios, see e.g. [7, 10, 14, 17, 19, 23, 31, 37]. Given the well-known pitfalls of DNNs with regard to a notorious uncertainty of optimization success, this may look disappointing at a first glance and offers little help with regard to (Q*). Nevertheless, an advantage is seen in the fact that a single approximation system can match the performance of other specialized methods in a diversity of scenarios, covering not only finite smoothness but the ability to simultaneously approximate fractal functions and holomorphic mappings very well, see e.g. [29, 37]. In addition it has also been shown that the actual approximation spaces for DNNs for algebraic approximation orders are significantly larger than the corresponding smoothness classes on which such an order has been established, [14].

In a different direction there are stunning results on *super-convergence* for DNN approximation on smoothness classes. Roughly speaking, when s is the degree of smoothness, the classical rate $O(n^{-s/d})$ (see (1.2)), can be doubled, i.e., DNNs achieve the rate $O(n^{-2s/d})$, $n \to \infty$, see [33, 35, 37]. Unfortunately, this is not enough to delineate DNNs from other approximation systems regarding the CoD.

Regarding new model classes, A. Barron's result in [1] has played a pioneering role for high dimensional approximation. The so called *greedy* construction of *shallow neural networks* with a single hidden layer [1] represents an instance of nonlinear approximation process that realizes dimension independent approximation rates when the model class is a *Barron class*, see e.g. [2, 34]. Note though that this model classes become "smaller" with increasing dimension. In fact, the defining conditions on the Fourier transform entail L_1-Besov smoothness of order $d/2$. Moreover, Barron spaces turn out to be approximation spaces for shallow (2-layer) networks, i.e., their elements

can be characterized through approximability by shallow networks at Monte Carlo rates, [12].

Barron spaces do, however, not characterize deep networks and are still subject to smoothness constraints. It should not come as a surprise that the expressive power of deep networks is closely tied to *compositions* of mappings, as DNNs themselves are special compositions. Several recent works are concerned with model classes based in different ways on compositional structures.

The authors in [27] characterize limits of compositions defined through sparsely connected graphs in the context of approximations on (high dimensional) Euclidean spheres. The important insight is the connection between the sparse graph-connectivity and the ability to avoid the CoD. This is also an important aspect in the approach to high dimensional nonlinear regression discussed in [32] which is actually closer in spirit to the present paper.

The authors in [12] propose so called compositional spaces as model classes to (nearly) characterize the approximation spaces for ReLU residual networks based on neural ODEs. As the authors mention themselves it is not so easy to see in a concrete application whether the target objects belong to these composition spaces. While these spaces describe certain limits of ReLU ResNet networks it is less clear how this pertains to the CoD.

In the present paper we propose a different approach to compositional model classes as described next.

1.3 Contributions and Layout

Our findings in response to the questions (Q), (Q*) revolve around two different yet intertwined main contributions:

(I) We propose and analyze a new notion of *compositional approximation spaces* to serve as model classes for high dimensional operator learning. They represent a certain "structural sparsity" or "regularity" in a broad sense that can help to avoid or mitigate the CoD, see Sect. 2.

(II) We apply these concepts to characterize the complexity of solution manifolds of parameter dependent *linear transport equations* (see Sect. 3) and show that the CoD can indeed be avoided when approximating corresponding parameter-to-solution maps.

We proceed putting (I) and (II) into proper perspective. Regarding (I), the perhaps main difference from previous works is that the regularity notion under (I) describes closeness of a given function to compositions with certain desirable *structural* as well as *stability* properties, see Sect. 2.4. We stress that these objects are still compositions of (continuous) functions that are not determined yet by a finite number of degrees of freedom. There is also no reference to specific neural network architectures or activation functions. In brief, the elements of the approximation classes are regularized limits of such *tamed compositions*, see Sect. 2.

Computational approximations take place at a second stage where the structural properties of the continuous objects can be exploited to arrive at a finitely parametrized DNN approximations, see Theorem 2.16 in Sect. 2.6. This approach is somewhat

motivated by [6] where tensor approximation spaces are shown to overcome the CoD when approximating solution manifolds of high dimensional diffusion equations in the following sense. It is shown there that a certain rate of "tensor-approximability" of the right hand side implies a "certain tensor approximability" of the solutions which can be viewed as a *regularity theorem* in a broad sense. Analogous results for DNN approximation form a further central objective in the present paper.

The specific taming conditions on compositions in the present context can to some extent be motivated through the celebrated Arnold–Kolmogorov Superposition Theorem, Sect. 2.2.

Although the primary application is to solution manifolds of transport equations, the reason for considering first compositions on a continuous level is to potentially widen the range of applicability. In general, elements of solution manifolds, as solutions to operator equations, can often be described in a constructive way as limits of iterative processes which naturally tie into limits of compositional structures, see also Sect. 5 for further comments.

Concerning (II), our interest in transport equations stems from the fact that they are a prototype of PDE models for which established methods for model reduction like Reduced Basis Methods (RBMs) fail. As explained above in Sect. 1.1, this is in stark contrast to *elliptic models* where the Kolmogorov n-widths decay robustly with respect to the parametric dimension.

Instead, for linear transport equations holomorphy of parameter-to-solution maps can only be established under very restrictive (and unrealistic) conditions on the convection field and the domain, see [20]. This appears to call for strictly nonlinear model reduction approaches and, to our knowledge, the authors in [24] are the first to explore DNN approximation for parametric transport equations. One should note though that the convergence rates established there still reflect the full CoD. As functions of spatial and parametric variables the constructed DNNs $\mathcal{N}_\varepsilon$ satisfy

$$\|u - \mathcal{N}_\varepsilon\|_{L_\infty} \leq \varepsilon, \quad \#\mathcal{N}_\varepsilon \eqsim \varepsilon^{-\frac{m+1+d_y}{\alpha}}, \quad \varepsilon \to 0, \tag{1.4}$$

where $m+1$, d_y denote the number of space-time and parametric variables and α represents the smoothness of the solutions resulting from suitable smoothness assumptions on the problem data (initial conditions, right hand sides, convection coefficients). The reason is that the authors assume *only* a certain degree of smoothness of the problem data which, in view of the comments in Sect. 1.1, the results are in essence best possible.

This indicates, in particular, that all one can expect is to leverage the properties of the PDE to derive a "regularity theorem" of the type mentioned above for tensor approximation in [6]. In brief: a certain regularity of the problem data—in the present situation, dimension sparse compositional approximability—implies a certain related degree of regularity (in the same sense) of the solutions.

Results of this type, concerning (II), are presented in Sect. 4. Recalling the way how solutions to transport equations depend on *characteristics*, the key to understanding compositional approximability of solutions is to quantify first compositional approximability of characteristics which is the subject of Sect. 4.1. For instance, when the

parameter dependence of the convection field is affine (a simple instance of dimension-sparse compositional approximability), the characteristics, as functions of m spatial and d_y parametric variables, can be approximated by DNNs as follows

$$\|z - \mathcal{N}_\varepsilon\|_{L_\infty([0,T]\times D\times\mathcal{Y};\mathbb{R}^m)} \le \varepsilon, \quad \#\mathcal{N}_\varepsilon \lesssim d_y\Big(\frac{e^{LT}}{\varepsilon}\Big)^{m+1}\Big|\log_2\frac{e^{LT}}{\varepsilon}\Big|^2,$$

avoiding, in contrast to (1.4), any exponential dependence on the large dimension d_y.

Based on these results, we present in Sect. 4.2 similar results that bound the complexity of *solution operators* as mappings on parameter dependent convection fields, initial data, and right hand sides, avoiding the CoD.

In Sect. 5 we close with indicating several directions of future research suggested by the present findings and their bearing on a wider problem scope.

All proofs are deferred to Sect. 6.

Notational Conventions: In what follows we often write $a \lesssim b$ to indicate that a is bounded by a constant multiple of b where the constant is independent of any parameters a and b may depend on, unless specified otherwise. Accordingly $a \eqsim b$ means $a \lesssim b$ and $b \lesssim a$.

For notational brevity and convenience we will use, for any pair of finite dimensional metric spaces $\mathbb{X}, \mathbb{Y}$ and any continuous function $g : \mathbb{X} \subset \mathbb{R}^{d_0} \to \mathbb{Y} \subset \mathbb{R}^{d_1}$, the shorthand notation

$$\|g\|_{\mathbb{X}} = \|g\|_{L_\infty(\mathbb{X};\mathbb{Y})} = \sup_{x\in\mathbb{X}}\sup_{i=1,\dots,d_1}|g_i(x)| = \sup_{x\in\mathbb{X}}|g(x)|_\infty,$$

when the particular domains and ranges don't matter.

Likewise we use the domain- and dimension independent notation $\|g\|_{\mathrm{Lip}_1}, |\cdot|_{\mathrm{Lip}_1}$ to denote the full Lipschitz norm, respectively semi-norm

$$|g|_{\mathrm{Lip}_1} := \sup_{x,z\in X}\frac{|g(x)-g(z)|}{|x-z|_\infty}, \quad \|g\|_{\mathrm{Lip}_1} := \max\{\|g\|_\infty, |g|_{\mathrm{Lip}}\}.$$

Our default meaning of $|\cdot|$ for vector-valued arguments is the max-norm.

We consistently denote generic scalar- or vector-valued functions by v. Composition factors are denoted by g, h, their compositions by G (which could be scalar- or vector-valued), spatial variables in $\mathbb{R}^m$, $m \in \{1, 2, 3\}$ by $x, z, w \in \mathbb{R}^m$. In particular, $z = z(\cdot,\cdot)$ as a function of time and space denotes a field of characteristics in $\mathbb{R}^m$. We reserve u to denote the solution of a PDE while f stands for corresponding right hand side. We use superscripts to index vector-valued quantities while subscripts enumerate their components.

2 Compositions

2.1 Compositional Representations

We will interpret "compositions"—denoted in what follows by the symbol "$\circ$" in $(g \circ h)(z) := g(h(z))$, in a broad sense, covering iterated applications of global operators as well as pointwise compositions of continuous functions. Here dimensional compatibility is always implicity assumed, i.e., $\dim \operatorname{range} h = \dim \operatorname{dom} g$. Specifically, we consider compositions of mappings

$$g^1 : D_0 \subset \mathbb{R}^{d_0} \to \mathbb{R}^{d_1}, \quad g^\ell : \mathbb{R}^{d_{\ell-1}} \to \mathbb{R}^{d_\ell}, \quad \ell = 2, \ldots, n,$$

where we always require that the last factor is linear, i.e., for some $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, d_{n-1}$

$$g^n = \sum_{i=1}^{d_{n-1}} \alpha_i g_i^{n-1}.$$

It will be convenient to abbreviate the ordered array of such dimension compatible mappings by $\mathbf{g} = (g^j)_{j=1}^n$ to provide a particular *realization*

$$G(z) = (g^n \circ \cdots \circ g^1)(z) =: G_{\mathbf{g}}(z), \tag{2.1}$$

of a mapping from $D_0 \subset \mathbb{R}^{d_0} \to \mathbb{R}^{d_n}$. We sometimes simply identify $\mathbf{g}$ with $G_{\mathbf{g}}$. A prominent example of compositional representations are DNNs whose formal definition can be found in numerous texts, see e.g. [7, 10, 15, 19, 29, 37]. Here we are content with mentioning that DNN realizations (denoted by $\mathcal{N}$) are (in their simplest feed-forward version) of the form (2.1) with factors $g^j(\cdot) = \sigma(A^j \cdot + b^j)$ where $A^j \in \mathbb{R}^{d_j \times d_{j-1}}$, $b^j \in \mathbb{R}^{d_j}$, and the *activation function* or *rectifier* σ acts componentwise. An important example is the ReLU rectifier $\sigma(t) = \max\{0, t\} =: t_+$. The entries in A^j, b^j or in the linear output layer are called *weights* and their number $\#\mathcal{N}$ is the *size* or *complexity* of the DNN $\mathcal{N}$. There are numerous important "architectural variants"like ResNet structures. We will address those as well as slight generalizations later below when the need arises.

2.2 The Arnold–Kolmogorov Superposition Theorem

The role of compositions, albeit not being emphasized as such, made an early appearance in the celebrated Arnold–Kolmogorov Superposition Theorem. In a variant established by G. G. Lorentz [26], it states that every continuous function G on $D_0 = [0, 1]^d$ has an *exact representation*

$$G(x_1, \ldots, x_d) = \sum_{q=0}^{2d} \Phi\Big(\sum_{p=1}^{d} \phi_{q,p}(x_p) \Big),$$

where Φ, $\phi_{q,p}$ are continuous functions. Thus, $G = g^4 \circ g^3 \circ g^2 \circ g^1$, where $g^1(x) = (\phi_{q,p}(x_p))_{q,p=0,1}^{2d,d}$, i.e., $d_1 = (2d+1)d$, $g^2(z) = \left(\sum_{p=1}^{d} z_{q,p}\right)_{q=0}^{2d}$, i.e., $d_2 = 2d+1$, $g^3(z) = (\Phi(z_q))_{q=0}^{2d}$, $d_3 = 2d+1$, $g^4(z) = \sum_{q=0}^{2d} z_q$. Hence, *every* $G \in C([0,1]^d)$ has a compositional representation of depth at most $n = 4$. Moreover, the structure of this composition is an instance of what is defined below as a 1-*dimension sparse* representation because composition factors are either affine or its components depend only on a single variable.

In principle, approximations to the univariate functions $\phi_{q,p}$ and an approximation to the univariate function Φ would produce an approximation to the high dimensional function G. While the $\phi_{p,q}$ can be made Lipschitz continuous, one can unfortunately not assert any additional smoothness of Φ beyond just continuity, even when G is assumed to have some positive degree of smoothness. So, realizing any target approximation accuracy for Φ could be arbitrarily expensive. This limits the direct use of the Superposition Theorem for practical purposes, see [13, 22] for somewhat controversial views in this regard.

From a positive perspective, as soon as Φ has some extra smoothness one would be able to defeat the CoD since only approximations to functions of a single variable would be required. This shows that dimension sparsity is in some sense a key property but by itself not sufficient to avoid the CoD.

The idea, underlying the following sparsity notion, is simply to relax "exact" to "approximable", allow for a larger depth, but keep the requirement that the compositional factors, unless having a simple explicit computable representation, depend only on a few variables, whose number stays much smaller than d. Moreover, to estimate accuracy of compositions one needs to quantify (at least a low degree of) smoothness of composition factors. These structural constraints give rise to what will be termed as *tamed compositions*.

2.3 Compositional s-Dimension Sparsity

Since

$$f \circ g = (f \circ h) \circ (h^{-1} \circ g) =: \tilde{f} \circ \tilde{g},$$

a mapping G may have infintely many compositional representations. In slight abuse of terminology we sometimes write $\mathbf{g}_G$ to express that $\mathbf{g}$ is a representation of the mapping G. Hence, one can always "reshape" compositional factors where the factors in the new represenation could have unfavorable regularity or stability properties, or in fact vice versa. Thus meaningful notions of compositional approximability require specifying some preferred structural properties. In the light of the discussion of the previous section, we impose two types of constraints:

(C1) We allow only representations $\mathbf{g}$ of a given G where all factors g^j are Lipschitz continuous.
(C2) We require the existence of a representation $\mathbf{g}$ which is *dimension sparse* in the sense described next.

We note that constraint (C2) draws some motivation also from the work on nonlinear regression in [32].

To that end, we quantify next the "cost" associated with a scalar valued function $v : \mathbb{R}^d \to \mathbb{R}$, in particular, taking the number of "active variables" into account. Define

$$\begin{gathered}\mathrm{s}(v) := 0, \quad \text{if } v(x_1, \dots, x_d) = x_i \text{ for some } i, \quad \mathrm{s}(v) := 1, \quad \text{if } v \text{ is multi-linear,} \\ \text{otherwise} \\ \mathrm{s}(v) := \min_{\mathrm{s} \le d} \{\exists\, \mathcal{I} \subset \{1, \dots, d\} : \#\mathcal{I} = \mathrm{s},\ \forall\, j \notin \mathcal{I},\ \ v(x_j : j \notin \mathcal{I}) \text{ is a constant}\}.\end{gathered} \tag{2.2}$$

In other words, we don't charge any cost to v if it just reproduces the value of one of its input-variables. It is charged one cost unit if it already has a finitely parametrized explicit represenation. In all other cases the cost attached to v is the number of variables it explicitly depends on.

Then, given a dimension compatible ordered array $\mathbf{g} = (g^j)_{j=1}^n$ let

$$\mathrm{s}_\infty(\mathbf{g}) := \max_{j \le n} \max_{i \le d_j} \mathrm{s}(g_i^j).$$

Definition 2.1 A representation $\mathbf{g}_G$ of a mapping G of the form (2.1) is called s-*dimension sparse* if $\mathrm{s}_\infty(\mathbf{g}) \le \mathrm{s}$.

This gives rise to the following measure for the *compositional complexity* of a representation $\mathbf{g}$

$$\mathfrak{N}(\mathbf{g}) := \sum_{j=1}^{n} \sum_{i=1}^{d_j} \mathrm{s}(g_i^j) \le \sum_{j=1}^{n} d_{j-1} d_j.$$

Remark 2.2 Thus, when $\mathbf{g}$ represents a DNN $\mathcal{N}$ dimension-sparsity corresponds to sparse connectivity and $\mathfrak{N}(\mathbf{g}) \eqsim \#\mathcal{N}$ which will be frequently used in what follows.

We say that two representations $\mathbf{g}, \mathbf{g}'$ are dimensionally compatible (in this order) if $G_{\mathbf{g}}, G_{\mathbf{g}'}$ are, i.e., if the output dimension $d_{n(\mathbf{g})}$ agrees with the input-dimension $d_0(\mathbf{g}')$.

Remark 2.3 For any two dimensionally compatible representations $\mathbf{g}, \mathbf{g}'$, the composition $\hat{G} := G_{\mathbf{g}'} \circ G_{\mathbf{g}}$ has a representation $\hat{\mathbf{g}} = (\mathbf{g}'|\mathbf{g})$ satisfying

$$\mathfrak{N}(\hat{\mathbf{g}}) = \mathfrak{N}(\mathbf{g}) + \mathfrak{N}(\mathbf{g}'). \tag{2.3}$$

Likewise, when $\mathbf{g}, \mathbf{g}'$ have equal in- and output dimension, i.e., $d_0(\mathbf{g}) = d_0(\mathbf{g}')$ and $d_{n(\mathbf{g})} = d_{n(\mathbf{g}')}$, the sum $\hat{G} := G_{\mathbf{g}} + G_{\mathbf{g}'}$ has a representation $\hat{\mathbf{g}} = \binom{\mathbf{g}}{\mathbf{g}'}$ satisfying

$$\mathfrak{N}(\hat{\mathbf{g}}) = \mathfrak{N}(\mathbf{g}) + \mathfrak{N}(\mathbf{g}'). \tag{2.4}$$

This follows easily by (what in the DNN context is called) parallelization (as hinted at by the notation $\binom{\mathbf{g}}{\mathbf{g}'}$), upon possibly inserting identity factors in the representation of smaller depth.

Given any integer s,

$$\mathfrak{C}_{N,\mathrm{s}} := \left\{ G : \exists\, \mathbf{g}_G \text{ s.t. } \mathfrak{N}(\mathbf{g}) \le N,\ \mathrm{s}_\infty(\mathbf{g}) \le \mathrm{s} \right\}$$

denotes then the collection of mappings with s-dimension-sparse representations of complexity at most N. When the input-domain $D \subset \mathbb{R}^d$ and output-dimension d' matters we write $\mathfrak{C}_{N,\mathrm{s}}(D, d')$.

$\mathfrak{C}_{N,\mathrm{s}}$ is of course not a linear set but, by Remark 2.3, one has (for compatible in- and output dimensions)

$$\mathfrak{C}_{N,\mathrm{s}} + \mathfrak{C}_{N',\mathrm{s}} \subset \mathfrak{C}_{N+N',\mathrm{s}}. \tag{2.5}$$

Likewise for mappings $G \in \mathfrak{C}_{N,\mathrm{s}}(D, d')$, $\tilde{G} \in \mathfrak{C}_{N',\mathrm{s}}(\mathbb{R}^{d'}, d'')$ we infer from (2.3) that

$$\tilde{G} \circ G \in \mathfrak{C}_{N+N',\mathrm{s}}(D, d''). \tag{2.6}$$

The following remarks motivate the discussion in the next section.

Remark 2.4 When $\mathrm{s} \ge d_0$, $D \subset \mathbb{R}^{d_0}$, the constraint of s-dimension sparsity is, of course void. In this case one simply has that $\mathfrak{C}_{N,\mathrm{s}}(D, d') = \mathrm{Lip}_1(D; \mathbb{R}^{d'})$. In fact, any $G \in \mathrm{Lip}_1(D, d')$ has a trivial representation

$$G(x) = (g^2 \circ g^1)(x), \quad g^1_i = G_i,\ i = 1, \ldots, d' \quad g^2 = \mathrm{id}_{d'}.$$

So, s-dimension-sparsity with $\mathrm{s} < d = d_0$ is essential for such a framework to offer interesting information.

2.4 Tamed Compositions and Approximation Classes

Since the classes $\mathfrak{C}_{N,\mathrm{s}}$ need not be closed, approximation by elements in $\mathfrak{C}_{N,\mathrm{s}}$ needs to be *regularized*. Assume that $\mathcal{R} : \mathbf{g} \mapsto \mathcal{R}(\mathbf{g}) \in \mathbb{R}_+$ complies with addition and composition in the sense that

$$\begin{aligned} \mathcal{R}\begin{pmatrix} \mathbf{g} \\ \mathbf{g}' \end{pmatrix} &\le \max\{\mathcal{R}(\mathbf{g}), \mathcal{R}(\mathbf{g}')\}, \quad \mathcal{R}(G_{\mathbf{g}} \circ G_{\mathbf{g}'}) \\ &\le \max\left\{\mathcal{R}(\mathbf{g}), \mathcal{R}(\mathbf{g}'), \mathcal{R}(\mathbf{g}) \cdot \mathcal{R}(\mathbf{g}')\right\}, \end{aligned} \tag{2.7}$$

where we assume dimensional compatibility in the second relation. We discuss instances of $\mathcal{R}$, based on (C1), later below.

Then for any $G \in \mathfrak{C}_{N,\mathrm{s}}$ let

$$|\!|\!| G |\!|\!|_{N,\mathrm{s},\mathcal{R}} = |\!|\!| G |\!|\!|_{N,\mathrm{s}} := \inf\left\{ \mathcal{R}(\mathbf{g}) : G_{\mathbf{g}} = G,\ \mathrm{s}(\mathbf{g}) \le \mathrm{s},\ \mathfrak{N}(\mathbf{g}) \le N \right\}. \tag{2.8}$$

We supress reference to $\mathcal{R}$ when this is clear from the context. We refer to $|\!|\!| \cdot |\!|\!|_{N,\mathrm{s},\mathcal{R}}$ as *"compositional norm"* although it is not a norm but close to one. In fact, the following relations follow from Remark 2.3, (2.5) and (2.6), combined with (2.7).

Remark 2.5 For any $G \in \mathfrak{C}_{N,\mathrm{s}}$ and $\tilde{G} \in \mathfrak{C}_{\tilde{N},\mathrm{s}}$ with the same in- and output dimensions one has

$$\|\!|G + \tilde{G}|\!\|_{N+\tilde{N},\mathrm{s}} \le \max\{\|\!|G|\!\|_{N,\mathrm{s}}, \|\!|\tilde{G}|\!\|_{\tilde{N},\mathrm{s}}\}. \tag{2.9}$$

Similarly, for dimensionally compatible mappings $G_i \in \mathfrak{C}_{N_i,\mathrm{s}}$, $i = 1, 2$, one has

$$\|\!|G_2 \circ G_1|\!\|_{N_1+N_2,\mathrm{s}} \le \max\Big\{\|\!|G_1|\!\|_{N_1,\mathrm{s}}, \|\!|G_2|\!\|_{N_2,\mathrm{s}}, \|\!|G_1|\!\|_{N_1,\mathrm{s}} \cdot \|\!|G_2|\!\|_{N_2,\mathrm{s}}\Big\}. \tag{2.10}$$

In the spirit of [6], consider the "K-functional"

$$K_{\mathrm{s}}(v, N, \delta) := \inf_{G \in \mathfrak{C}_{N,\mathrm{s}}} \|v - G\|_{L_\infty} + \delta \|\!|G|\!\|_{N,\mathrm{s}}. \tag{2.11}$$

Obviously, $K_{\mathrm{s}}(v, N, \delta) \le C\delta$ means that for some $\bar{G} \in \mathfrak{C}_{N,\mathrm{s}}$ one has $\|v - \bar{G}\|_{L_\infty} \le C\delta$ and $\|\!|\bar{G}|\!\|_{N,\mathrm{s}} \le C$, i.e., accuracy δ is achieved with a controlled "composition-norm".

Interrelating N and δ is then a way to define collections of functions with a certain quantifiable *compositional approximability*. A (smooth) strictly increasing function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{s\to\infty} \gamma(s) = \infty$, is called a *growth function*. Its inverse γ^{-1} exists and is also a growth function. This is to be distinguished from $\gamma(\cdot)^{-1} = 1/\gamma(\cdot)$. Given such a growth function γ, we consider the *compositional approximation class*

$$\mathcal{A}^{\gamma,\mathrm{s}} := \{v \in \mathbb{X} : \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}} := \|v\|_\infty + |v|_{\mathcal{A}^{\gamma,\mathrm{s}}} < \infty\},$$

where $|v|_{\mathcal{A}^{\gamma,\mathrm{s}}} := \sup_{N\in\mathbb{N}} \gamma(N) K_{\mathrm{s}}(v, N, \gamma(N)^{-1})$.

Since trivially

$$\|v - G_N\|_{L_\infty} \le \gamma(N)^{-1}\gamma(N)\big\{\|v - G_N\|_{L_\infty} + \gamma(N)^{-1}\|\!|G_N|\!\|_{N,\mathrm{s}}\big\} \le \gamma(N)^{-1}|v|_{\mathcal{A}^{\gamma,\mathrm{s}}},$$

we will use the information $v \in \mathcal{A}^{\gamma,\mathrm{s}}$ often in the form that for each $N \in \mathbb{N}$ there exists $G_N \in \mathfrak{C}_{N,\mathrm{s}}$ such that

$$\|v - G_N\|_{L_\infty} \le \gamma(N)^{-1}|v|_{\mathcal{A}^{\gamma,\mathrm{s}}}, \quad \|\!|G_N|\!\|_{N,\mathrm{s}} \le \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}}, \quad N \in \mathbb{N}. \tag{2.12}$$

To put it differently, realizing target accuracy ε by a composition $G_{\mathbf{g}_\varepsilon}$ is achievable within a complexity $N = N_\varepsilon$ of the order

$$N_\varepsilon = \left\lceil \gamma^{-1}\Big(\frac{|v|_{\mathcal{A}^{\gamma,\mathrm{s}}}}{\varepsilon}\Big)\right\rceil. \tag{2.13}$$

Remark 2.6 Finer scales $\mathcal{A}^{\gamma,\beta,\mathrm{s}}$, $\beta \in (0, 1]$, of approximation classes can be obtained by defining $|v|_{\mathcal{A}^{\gamma,\beta,\mathrm{s}}} := \sup_{n\in\mathbb{N}} \gamma(N)^\beta K_{\mathrm{s}}(v, N, \gamma(N)^{-1})$. This implies the existence of $G_N \in \mathfrak{C}_{N,\mathrm{s}}$ such that $\|v - G_N\|_{L_\infty} \le \gamma(N)^{-\beta}|v|_{\mathcal{A}^{\gamma,\beta,\mathrm{s}}}$, while $\gamma(N)^{\beta-1}\|\!|G_N|\!\|_{N,\mathrm{s}} \le \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}}$. So, for $\beta < 1$ the composition norms are allowed to grow like at most $\gamma(N)^{1-\beta}$, $N \in \mathbb{N}$. Since in subsequent applications a uniform control on composition norms matter, we confine the discussion henceforth to the special case $\beta = 1$

2.5 Choices for $\mathcal{R}$ and Basic Properties of $\mathfrak{C}_{N,\mathrm{s}}$ and $\mathcal{A}^{\gamma,\mathrm{s}}$

Although other variants of $\mathcal{R}$ are conceivable we focus here on

$$\mathcal{R}(\mathbf{g}) := \max\left\{\|g^\ell\|_{\mathrm{Lip}_1}, L_{[n(\mathbf{g}),\ell+1]}(\mathbf{g}) : \ell = 1,\dots,n(\mathbf{g})\right\}, \tag{2.14}$$

where $L_{[n(\mathbf{g}),\ell+1]}(\mathbf{g})$ denotes the Lipschitz constant of the *partial compositions* $g^{n(\mathbf{g})} \circ \cdots \circ g^{\ell+1}$. One easily verifies that $\mathcal{R}$ satisfies (2.7). Controlling $|\!|\!| \cdot |\!|\!|_{N,\mathrm{s}}$, obviously constrains representations of elements further. In particular, for $\mathcal{R}$, given by (2.14),

$$\|G\|_{\mathrm{Lip}_1} \le |\!|\!| G |\!|\!|_{N,\mathrm{s}}, \quad G \in \mathfrak{C}_{N,\mathrm{s}}. \tag{2.15}$$

Remark 2.7 It will be at times useful to consider alternate weaker regularizers. A natural alternative would be

$$\mathcal{R}^\circ(\mathbf{g}) := \max\left\{\|g^\ell\|_{\mathrm{Lip}_1} : \ell = 1,\dots,n(\mathbf{g})\right\}, \tag{2.16}$$

so that trivially

$$|\!|\!| G |\!|\!|_{N,\mathrm{s},\mathcal{R}^\circ} \le |\!|\!| G |\!|\!|_{N,\mathrm{s},\mathcal{R}} \le \max\{1, |\!|\!| G |\!|\!|^N_{N,\mathrm{s},\mathcal{R}^\circ}\}. \tag{2.17}$$

Remark 2.8 The definition of $\mathcal{A}^{\gamma,\mathrm{s}}$ makes sense for any regularizer satisfying (2.7). If we want to specify any other regularizer than the default one (2.14) we indicate this by a corresponding subscript such as e.g. $\mathcal{A}^{\gamma,\mathrm{s}}_{\mathcal{R}^\circ}$.

Another way of weakening $|\!|\!| \cdot |\!|\!|_{N,\mathrm{s}}$ is to replace the compositional Lipschitz constants $L_{[n(\mathbf{g}),\ell+1]}(\mathbf{g})$ in (2.14) by $\zeta(N)L_{[n(\mathbf{g}),\ell+1]}(\mathbf{g})$, where $\zeta(N) \to 0$. This would permit some growth of the $L_{[n(\mathbf{g}),\ell+1]}(\mathbf{g})$ and hence no longer forces compositions to stay in Lip_1. In view of the applications to come, we focus in what follows on the stronger version (2.14), which implies (2.15), see also Remark 2.6.

Proposition 2.9 *For any fixed constant $B < \infty$ and any $\mathrm{s} \in \mathbb{N}$ the set*

$$\mathfrak{C}_{N,\mathrm{s}}(B) := \{G \in \mathfrak{C}_{N,\mathrm{s}} : |\!|\!| G |\!|\!|_{N,\mathrm{s}} \le B\}$$

is compact in $L_\infty := L_\infty(D)$ for either regularizor $\mathcal{R}$ or $\mathcal{R}^\circ$. Hence, a minimizing representation in (2.8) exists.

The proof of Proposition 2.9 is given in Appendix A. Again, in view of Remark 2.4 and (2.15), compactness of $\mathfrak{C}_{N,\mathrm{s}}(B)$ is trivial when $\mathrm{s} \ge d_0$ or $\mathcal{R}$ is given by (2.14).

We briefly discuss some basic consequences for the approximation classes $\mathcal{A}^{\gamma,\mathrm{s}}$.

Again, on account of Remark 2.4 the classes $\mathcal{A}^{\gamma,\mathrm{s}}_{\mathcal{R}}$, $\mathcal{A}^{\gamma,\mathrm{s}}_{\mathcal{R}^\circ}$ agree with Lip_1 when $\mathrm{s} \ge d_0$ and hence do not provide any useful information. In particular, by (2.15) and (2.12), balls in these classes are compact when $\mathrm{s} \ge d_0$. Since $\|v\|_{\mathrm{Lip}_1} \le \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}_{\mathcal{R}}}$ for all $\mathrm{s} \in \mathbb{N}$, precompactness of $\mathcal{K}_{\gamma,\mathrm{s}}(B)$ is also immediate. Since in general $\mathcal{A}^{\gamma,\mathrm{s}}_{\mathcal{R}^\circ}$ need not be contained in Lip_1 for $\mathrm{s} < d$ the following claim requires an argument provided in Appendix A.

Remark 2.10 The unit ball

$$U\mathcal{A}^{\gamma,\mathrm{s}} := \{v \in \mathcal{A}^{\gamma,\mathrm{s}} : \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}} \leq 1\}$$

is for any $\mathrm{s} \in \mathbb{N}$ compact in $C(D)$.

We record a few elementary properties of the classes $\mathcal{A}^{\gamma,\mathrm{s}}$.

Remark 2.11

(a) Obviously one has

$$\gamma(\cdot) \lesssim \tilde{\gamma}(\cdot) \quad \Rightarrow \quad \mathcal{A}^{\tilde{\gamma},\mathrm{s}} \subseteq \mathcal{A}^{\gamma,\mathrm{s}}.$$

(b) Whenever γ satisfies $\gamma(N) \geq c_\gamma \gamma(2N)$ for $N \in \mathbb{N}$, one has $\mathcal{A}^{\gamma,\mathrm{s}} + \mathcal{A}^{\gamma,\mathrm{s}} \subset \mathcal{A}^{\gamma,\mathrm{s}}$.

Remark 2.12

(a) For each $N \in \mathbb{N}$ (fixed) $\mathfrak{C}_{N,\mathrm{s}} \subset \mathcal{A}^{\gamma,\mathrm{s}}$ for *every* growth function γ.
(b) Assume that $\gamma \lesssim \tilde{\gamma}$ and $\gamma(N) \geq c_\gamma \gamma(2N)$ for a fixed $c_\gamma > 0$. For respective equal input- and output-dimension and $\mathcal{R}$ according to (2.14), $\mathcal{A}^{\gamma,\mathrm{s}}$ is closed under composition with elements from $\mathcal{A}^{\tilde{\gamma},\mathrm{s}}$. An analogous statement holds when $v \in \mathcal{A}^{\gamma,\mathrm{s}}(D, d')$, $w \in \mathcal{A}^{\tilde{\gamma},\mathrm{s}}_{\mathcal{R}^\circ}(D', d'')$, $\dim D' = d'$. The growth range covers any polynomial growth.

Proof (a) is obvious.

Regarding (b), for $v \in \mathcal{A}^{\gamma,\mathrm{s}}$ and $w \in \mathcal{A}^{\tilde{\gamma},\mathrm{s}}$, $N \in \mathbb{N}$, there exist G_w, $G_v \in \mathfrak{C}_{N,\mathrm{s}}$ (with respective in- and output-dimensions), such that $\|v - G_v\|_{L_\infty} \leq \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}} \gamma(N)^{-1}$, and $\|w - G_w\|_{L_\infty} \leq \|w\|_{\mathcal{A}^{\tilde{\gamma},\mathrm{s}}} \tilde{\gamma}(N)^{-1}$. Thus, since by (2.6), $G_w \circ G_v \in \mathfrak{C}_{2N,\mathrm{s}}$, we use (2.15) to conclude

$$\begin{aligned}
\|w \circ v - G_w \circ G_v\|_{L_\infty} &\leq \|(w - G_w) \circ v\|_{L_\infty} + \|G_w(v) - G_w(G_v)\|_{L_\infty} \\
&\leq \|w\|_{\mathcal{A}^{\tilde{\gamma},\mathrm{s}}} \tilde{\gamma}(N)^{-1} + |G_w|_{\mathrm{Lip}_1} \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}} \gamma(N)^{-1} \\
&\leq c_\gamma^{-1} \|w\|_{\mathcal{A}^{\tilde{\gamma},\mathrm{s}}} \left\{1 + \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}}\right\} \gamma(2N)^{-1}.
\end{aligned}$$

Since by (2.10) and (2.12), $\|\!|G_w \circ G_v|\!\|_{2N,\mathrm{s}} \leq \max\left\{\|w\|_{\mathcal{A}^{\tilde{\gamma},\mathrm{s}}}, \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}}, \|w\|_{\mathcal{A}^{\tilde{\gamma},\mathrm{s}}} \cdot \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}}\right\}$, (b) follows. □

2.6 From Continuous to Discrete

2.6.1 Lipschitz-Stable Neural Network Approximation

The approximation classes introduced above characterize approximability by "tamed" or "regularized" compositions which themselves are not yet described by finitely many parameters. However, the compositional structure leads, in a second step, to a *finitely parametrized approximation*, see also [32] for an approach in the same spirit. The

relevant property is the first constraint (C1) and the control of Lipschitz norms enforced by the above choices of $\mathcal{R}$, see (2.14), (2.16), and their role in (2.12).

Specifically, perturbing a tamed composition, by approximating all or some of its components, (C1) allows one to estimate the overall error. The following is a simple folklore perturbation bound. Consider dimensionally compatible Lipschitz functions g, h with constants L_g, L_h and $\varepsilon_g, \varepsilon_h$ accurate approximations $\tilde{g}, \tilde{h}$. Then

$$\begin{aligned}\|g \circ h - \tilde{g} \circ \tilde{h}\|_{L_\infty} &\leq \|g \circ h - g \circ \tilde{h}\|_{L_\infty} + \|g \circ \tilde{h} - \tilde{g} \circ \tilde{h}\|_{L_\infty} \\ &\leq L_g \varepsilon_h + \varepsilon_g. \end{aligned} \tag{2.18}$$

Given a compositional representation $G_{\mathbf{g}}$ of the form (2.1), we denote for $k \leq n$ by $L_{[n,k]} = L_{[n,k]}(\mathbf{g})$ the Lipschitz constant of partial compositions $g^n \circ g^{n-1} \circ \cdots \circ g^k$. It will be convenient to set $L_{[n,n+1]} = 1$. Then, using the above argument inductively yields the following familiar facts. see e.g. [32].

Remark 2.13 Assume that we have mappings $\tilde{g}^j : \mathbb{R}^{d_{j-1}} \to \mathbb{R}^{d_j}$, $j = 1, \ldots, n$, such that

$$\|g^j - \tilde{g}^j\|_{L_\infty} \leq \varepsilon_j, \quad j = 1, \ldots, n.$$

Then

$$\|g^n \circ \cdots \circ g^1 - \tilde{g}^n \circ \cdots \circ \tilde{g}^1\|_{L_\infty} \leq \varepsilon_n + \sum_{j=1}^{n-1} \varepsilon_j L_{[n,j+1]} = \sum_{j=1}^{n} \varepsilon_j L_{[n,j+1]}.$$

By a symmetric argument we can replace $L_{[n,j+1]}(\mathbf{g})$ by $L_{[n,j+1]}(\tilde{\mathbf{g}})$.

In terms of the individual Lipschitz constants, one has, of course, $L_{[n,k]} \leq \prod_{j=k}^{n} L_j$. Finally, the estimates remain valid for more general iterated applications of operators as long as an estimate like (2.18) holds.

An important context where this will be used later is the following result.

Proposition 2.14 *For any $\delta > 0$, and any $v \in \mathrm{Lip}_1((0,1)^{\mathrm{S}})$, there exsist a ReLU network $\mathcal{N}_\delta$ such that*

$$\|v - \mathcal{N}_\delta\|_\infty \leq \delta, \quad \|\mathcal{N}_\delta\|_{\mathrm{Lip}_1} \leq c_3(1 + \|v\|_\infty)\|v\|_{\mathrm{Lip}_1}, \tag{2.19}$$

and

$$\#\mathcal{N}_\delta \leq c_1 \|v\|^{\mathrm{S}}_{\mathrm{Lip}_1} \delta^{-\mathrm{S}} \log_2 \frac{1}{\delta}, \quad \textit{depth of } \mathcal{N}_\delta \leq c_2 \log_2 \frac{1}{\delta}, \tag{2.20}$$

where the constants $c_1, c_2.c_3$ depend only on S. *Using if necessary Lipschitz-stable continuation from bounded domains to hypercubes, analogous results hold for more general domains under mild geometric constraints (see e.g. [19]).*

Gühring et al. [19] establishes the existence of ReLU networks that approximate functions of higher Sobolev regularity in weaker Sobolev norms *without* compromizing the standard complexity bounds. If we imposed more regularity than just $g \in \mathrm{Lip}_1$,

these results would imply (2.19) with a constant tending to one. In the present context we prefer to avoid assuming such "excess regularity" and sketch corresponding arguments in Appendix A for completeness, building on some of the concepts in [19].

Remark 2.15 Lipschitz functions of s variables are known to be approximable by deep networks at a *super-convergence rate*. That is, one achieves accuracy ε at the expense of $O(\varepsilon^{-s/2})$ rather than $O(\varepsilon^{-s})$ weights, see [33, 35, 37]. In the present context it will be important though to limit depth and to control in addition the Lipschitz stability of approximating networks. We refer to the discussion in [10] regarding the compatibility of super convergence and stable approximability. Besides, even the super-convergence rates would not avaoid the CoD. Therefore, we are content with the above non-optimal rates.

2.6.2 Implanting Finitely-Parametrized Components

"Implanting Lipschitz-stable neural networks" means that every component (non-trivial) g_i^j in a factor g^j with $s(g^j) > 1$ is replaced by a neural network (with in- and output-dimension one), so as to produce an expanded composition whose factors are either at most bilinear or neural networks. So in total they form a finitely parametrized function which we still refer to as a neural network. Note that this approximation does preserve s-dimension-sparsity.

The next result reflects the underlying guiding principle.

Theorem 2.16 *Assume that $v \in \mathcal{A}^{\gamma,s}$, then for every $\varepsilon > 0$ there exists a DNN $\mathcal{N}_\varepsilon$, such that*

$$\|v - \mathcal{N}_\varepsilon\|_{L_\infty} \leq \varepsilon, \tag{2.21}$$

where

$$\begin{aligned} \#\mathcal{N}_\varepsilon \lesssim\ & 2^s \Big(\frac{\|v\|^2_{\mathcal{A}^{\gamma,s}}}{\varepsilon}\Big)^s \Big(\gamma^{-1}\Big(\frac{2|v|_{\mathcal{A}^{\gamma,s}}}{\varepsilon}\Big)\Big)^{s+1} \\ & \max\Big\{\log_2 \frac{2\|v\|_{\mathcal{A}^{\gamma,s}}}{\varepsilon}, \log_2 \gamma^{-1}\Big(\frac{2|v|_{\mathcal{A}^{\gamma,s}}}{\varepsilon}\Big)\Big\}. \end{aligned} \tag{2.22}$$

Moreover, one has

$$|||\mathcal{N}_\varepsilon|||_{N_\varepsilon,s} \lesssim \|v\|^{N_\varepsilon}_{\mathcal{A}^{\gamma,s}}, \quad N_\varepsilon \eqsim \gamma^{-1}\big(2|v|_{\mathcal{A}^{\gamma,s}}/\varepsilon\big), \tag{2.23}$$

while for the weaker regularization $\mathcal{R}^\circ$ one has $|||\mathcal{N}_\varepsilon|||_{N_\varepsilon,s,\mathcal{R}^\circ} \lesssim \|v\|_{\mathcal{A}^{\gamma,s}_{\mathcal{R}}}$.

Thus, unless $\|v\|_{\mathcal{A}^{\gamma,s}}$ hides an exponential dependence on the large input dimension d_0, unit balls $U\mathcal{A}^{\gamma,s}$ are model classes for which DNN approximation avoids the Curse of Dimensionality when $s \ll d_0$.

It is instructive to specialize these estimates for two types of growth functions

$$\text{(alg):}\quad \gamma(r) \eqsim C_a r^\alpha, \quad \text{or} \qquad \text{(exp):}\quad \gamma(r) \eqsim C_e e^{\alpha r}, \tag{2.24}$$

for some $\alpha > 0$. (With a bit more technical effort the arguments extend to more refined scales like $\gamma(r) \sim e^{\alpha r^\beta}$, for some $0 < \beta \leq 1$.) For convenience we record for frequent future use

$$\gamma^{-1}(s) \eqsim \begin{cases} C_a^{-1/\alpha} s^{1/\alpha}, & \gamma \sim (\text{alg}),\ \alpha \geq \alpha_0 > 0; \\ \frac{1}{\alpha} \ln \frac{s}{C_e}, & \gamma \sim (\exp),\ \alpha > 0. \end{cases} \tag{2.25}$$

Hence (2.22) takes the form

$$\mathcal{N}_\varepsilon \lesssim \begin{cases} \|v\|^{\mathrm{s}}_{\mathcal{A}^\gamma,\mathrm{s}} \left(\frac{\|v\|_{\mathcal{A}^\gamma,\mathrm{s}}}{\varepsilon}\right)^{\frac{\mathrm{s}(\alpha+1)+1}{\alpha}} \left|\log_2 \frac{\|v\|_{\mathcal{A}^\gamma,\mathrm{s}}}{\varepsilon}\right|, & \gamma \sim (\text{alg}), \\ \left(\frac{\|v\|^2_{\mathcal{A}^\gamma,\mathrm{s}}}{\varepsilon}\right)^{\mathrm{s}} \left|\log_2 \frac{\|v\|_{\mathcal{A}^\gamma,\mathrm{s}}}{\varepsilon}\right|^{2+\mathrm{s}}, & \gamma \sim (\exp). \end{cases}$$

When $\gamma \sim (\exp)$ strong stability in (2.23) deteriorates only slowly according to $\|v\|^{|\ln \varepsilon|}_{\mathcal{A}^\gamma,\mathrm{s}}$.

In general, the stronger the algebraic growth order the closer the dominating complexity factor comes to the rate $\varepsilon^{-\mathrm{s}}$ which is attained for exponential growth (up to logarithmic factors). This rate is what one can expect for Lipschitz functions of s variables.

The assertion of Theorem 2.16 hinges on the following Lemma which we state here for later reference in several applications of similar type.

Lemma 2.1 *Assume that for some* $\mathrm{s} \leq d_0$, *the mapping* G *belongs to* $\mathfrak{C}_{N,\mathrm{s}}$, *i.e., is* s*-dimension-sparse (see Definition 2.1). Let* $G = G_{\mathbf{g}} \in \mathfrak{C}_{N,\mathrm{s}}$. *Let the DNN* $\mathcal{N}$ *be obtained by replacing each component* g_i^j *of each factor* g^j *with* $\mathrm{s}(g^j) > 1$, *by a* δ_j *accurate Lipschitz stable network* $\mathcal{N}_i^j$, *i.e.,*

$$\begin{aligned} &\|g_i^j - \mathcal{N}_i^j\|_{L_{\infty i}} \leq \delta_j, \\ &\|\mathcal{N}_j^i\|_{\mathrm{Lip}_1} \leq c_3(1 + \|g_i^j\|_\infty)\|g_i^j\|_{\mathrm{Lip}_1}, \quad i = 1, \ldots, d_j,\ j = 1, \ldots, n(\mathbf{g}) - 1. \end{aligned}$$

Then one has

$$\begin{aligned} \|G - \mathcal{N}\|_{L_\infty} &\leq \delta_n + \sum_{j=1}^{n(D)-1} \delta_j L_{[n(D),j+1]}, \\ \#\mathcal{N} &\leq \sum_{j=1}^{n(D)-1} \|g^j\|^{\mathrm{s}}_{\mathrm{Lip}} d_j \delta_j^{-\mathrm{s}} |\log_2 \delta_j|. \end{aligned} \tag{2.26}$$

and

$$\#\mathcal{N} \leq N \|G\|^{\mathrm{s}}_{N,\mathrm{s}} \max_{j=1,\ldots,n(D)} \delta_j^{-\mathrm{s}} |\log_2 \delta_j|. \tag{2.27}$$

Proof : The first relation in (2.26) follows from Remark 2.13 while the second one is a consequence of Proposition 2.14 together with Remark 2.2. Since $\sum_{j=1}^{n(\mathbf{g})} d_j \leq \mathfrak{N}(\mathbf{g})$, (2.27) follows from the definition of the compositional norms. □

We return to the proof of Theorem 2.16. By (2.12) we can find for each $N \in \mathbb{N}$ a $G_N \in \mathfrak{C}_{N,\mathrm{S}}$ satisfying (2.12). Given $\varepsilon > 0$, (2.13) says that $N_\varepsilon = \gamma^{-1}\big(2|v|_{\mathcal{A}^{\gamma,\mathrm{s}}}/\varepsilon\big)$ (we ignore the ceil-opertor) suffices to ensure that

$$\|v - G_{N_\varepsilon}\|_{L_\infty} \le \frac{\varepsilon}{2}.$$

Let $\mathcal{N}_{N_\varepsilon,\delta}$ denote the DNN, obtained by Lemma 2.1, with implant-tolerances $\delta = \delta_j$ all equal. to conclude that for any minimizing representation $\mathbf{g}_{N_\varepsilon}$ of G_{N_ε}

$$\|v - \mathcal{N}_{N_\varepsilon,\delta}\|_{L_\infty} \le \frac{\varepsilon}{2} + \delta n(\mathsf{D}(\mathbf{g}_{N_\varepsilon}))|\!|\!|G_{N_\varepsilon}|\!|\!|_{N_\varepsilon,\mathrm{S}} \le \frac{\varepsilon}{2} + N_\varepsilon \delta \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}}.$$

Choosing $\delta = \delta(\varepsilon) := \varepsilon/(2N_\varepsilon\|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}})$, produces a network $\mathcal{N}_\varepsilon := \mathcal{N}_{N_\varepsilon,\delta(\varepsilon)}$ satisfying (2.21). Regarding (2.22), we infer now from $|\!|\!|G_{N_\varepsilon}|\!|\!|_{N_\varepsilon,\mathrm{S}} \le \|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}}$ and (2.27) that

$$\#\mathcal{N}_\varepsilon \le N_\varepsilon \|v\|^{\mathrm{S}}_{\mathcal{A}^{\gamma,\mathrm{s}}} \delta(\varepsilon)^{-\mathrm{S}} |\log_2 \delta(\varepsilon)| = 2^{\mathrm{S}} \|v\|^{2\mathrm{S}}_{\mathcal{A}^{\gamma,\mathrm{s}}} \varepsilon^{-\mathrm{S}} N_\varepsilon^{1+\mathrm{S}} \Big|\log_2 \frac{2N_\varepsilon\|v\|^{\mathrm{S}}_{\mathcal{A}^{\gamma,\mathrm{s}}}}{\varepsilon}\Big|.$$

Since $N_\varepsilon = \gamma^{-1}\big(2|v|_{\mathcal{A}^{\gamma,\mathrm{s}}}/\varepsilon\big)$, the assertion (2.22) follows.

Finally, regarding the stability of the networks $\mathcal{N}_\varepsilon$, we employ the (possibly over-pessimistic) estimate (2.17) to obtain (2.23). □

Remark 2.17 Time-stepping in discretized dynamical systems is not the only context where one can expect to encounter compositional sparsity. More generally, solutions to operator equations can often be approximated by *iterative processes* such as fixed-point iterations that may help to assert membership to a compositional approximation class. This is exemplified next for a specific scenario where standard reduced modeling concepts suffer from slowly decaying Kolmogorov widths.

3 Linear Parametric Transport Equations

3.1 A Model Problem

We consider the Cauchy problem for a linear (scalar) transport equation in m spatial dimensions ($m \in \{1, 2, 3\}$, say) with parameter dependent data

$$\begin{aligned} \partial_t u(t,x) + \mathbf{a}(t,x,y)\cdot\nabla_x u(t,x) - f(t,x,y) = 0, \quad x \in \mathbb{R}^m,\ t \in [0,\widehat{T}),\ y \in \mathcal{Y},\\ u(0,x,y) = u_0(x,y), \quad x \in \mathbb{R}^m,\ y \in \mathcal{Y}, \end{aligned} \tag{3.1}$$

which is a standard format for models with uncertain data. We assume for convenience that $\operatorname{supp} u_0 = \overline{D}\times\mathcal{Y}$ where D is a bounded domain. Hence, for a fixed time horizon $\widehat{T}$ the solution, as a function of t, x can take values different from zero only in a bounded subset of $[0,\widehat{T})\times\mathbb{R}^m$. In what follows $\widehat{T}$ should be viewed as fixed finite but possibly

large whose order of magnitude is expected to affect the complexity of the envisaged parameter-to-solution maps $y \mapsto u(y)$.

We shall sometimes view $u(y)$ for each $y \in \mathcal{Y}$ as a function of $(t, x) \in [0, \widehat{T}) \times \mathbb{R}^m$, i.e., as a "point" in $L_\infty([0, \widehat{T}) \times \mathbb{R}^m)$.

To generate in the end efficient surrogates for the parameter-to-solution map, it will nevertheless be useful to view u as a function of all variables (t, x, y)

$$u : \Omega := [0, \widehat{T}) \times \mathbb{R}^m \times \mathcal{Y} \to \mathbb{R},$$

so that for *high parameter-dimensionality* $\mathcal{Y} \subset \mathbb{R}^{d_y}$, $d_y \gg 1$, one faces approximation problems in high dimensions. Since we are interested in conditions other than smoothness that may help avoiding the Curse of Dimensionality we impose only low or moderate smoothness conditions on the problem data. Specifically, we assume throughout:

$$\begin{gathered}\mathbf{a} \in C(0, \widehat{T}; \mathrm{Lip}_1(\mathbb{R}^m \times \mathcal{Y})), \quad \|\mathbf{a}\|_{L_\infty(\Omega;\mathbb{R}^m)} \le A, \\ |\mathbf{a}(t, z; y) - \mathbf{a}(t, z', y')| \le L \max\{|z - z'|, |y - y'|\}, \quad (t, z, y), (t, z', y') \in \Omega.\end{gathered} \tag{3.2}$$

In addition we require at times Lipschitz-in-time continuity of $\mathbf{a}$

$$\mathbf{a} \in \mathrm{Lip}_1(0, \widehat{T}; C(\mathbb{R}^m \times \mathcal{Y})), \quad |\mathbf{a}(\cdot, w)|_{\mathrm{Lip}_1([0,\widehat{T}])} \le L_t. \tag{3.3}$$

We separate the Lipschitz-conditions (3.2) and (3.3) because (3.3) is, under certain circumstances not necessary, see Remark 3.1 in the next section.

To see what one can expect regarding sparsity of solutions, the very special case, where $\mathbf{a}$ is independent of t, x, is instructive. E.g. when $f = 0$ the solution $u(t, x, y) = u_0(x - t\mathbf{a}(y), y)$ is a simple *composition* of u_0 with a linear function in (t, x) involving, however, a y-dependent coefficient. Even when u_0 does not depend on y but $\mathbf{a}(y)$ can be any element in $\mathrm{Lip}_1(\mathcal{Y})$ the solution can be in essence an arbitrary Lipschitz function and, as pointed out in Sect. 1.1, stable approximations will suffer from the CoD. The same holds, if $\mathbf{a}(y) = \mathbf{a}$ is constant but $u_0(\cdot, y)$ is an arbitrary element in a $\mathrm{Lip}_1(\mathcal{Y})$-ball. Analogous considerations apply to the right hand side f when only smoothness conditions are imposed. This is in agreement with the findings in [24] where the only conditions on the convection field are given in terms of classical smoothness properties.

In conclusion, more specific structural constraints on the data are needed to ensure that u can be approximated without suffering from the Curse. In brief, all one can expect is a "heredity" effect where some structural sparsity of the data leads to a structural solution sparsity that allows one to avoid the CoD.

3.2 Dimension-Sparse Compositional Convection Fields

In the light of the preceding comments we consider convection fields $\mathbf{a}$ that belong to the Bochner-type space of functions that are continuous in time with uniformly controlled values in $\mathcal{A}^{\gamma,\mathrm{S}} = \mathcal{A}^{\gamma,\mathrm{S}}_{\mathcal{R}}$, $\mathcal{R}$ given by (2.14)

$$\mathbf{a} \in L_\infty([0, \widehat{T}); \mathcal{A}^{\gamma,\mathrm{S}}(\mathbb{R}^m \times \mathcal{Y}; \mathbb{R}^m)). \tag{3.4}$$

The regularization (2.14) is used in the definition of $\mathcal{A}^{\gamma,\mathrm{s}}$ because, under the above assumptions, solutions and characteristics belong to $\mathrm{Lip}_1(\Omega)$.

Here and below $m \le \mathrm{s} \le m+1+d_y$ marks some dimension-sparsity with respect to the total number of variables. The fact that we do not assume $\mathbf{a} \in \mathcal{A}^{\gamma,\mathrm{s}}(\Omega;\mathbb{R}^m)$ indicates that the time variable receives a special treatment.

Remark 3.1 Time-Lipschitz continuity (3.3) is not always necessary. For our purposes it would suffice to know that compositional approximability is inherited by time-averages $\mathbf{a}_I(\bar{z};y) := |I|^{-1}\int_I \mathbf{a}(s,\bar{z};y)ds \in \mathcal{A}^{\gamma,\mathrm{s}}$, i.e.,

$$\|\mathbf{a}_I\|_{\mathcal{A}^{\gamma,\mathrm{s}}} \lesssim \|\mathbf{a}\|_{L_\infty(0,\widehat{T};\mathcal{A}^{\gamma,\mathrm{s}})}, \quad \forall\ I \subset [0,\widehat{T}]. \tag{3.5}$$

This condition holds e.g. in the case of affine parameter dependence, introduced next, or when $\mathbf{a}(t,x;y) = \sum_{i=1}^r \mathbf{a}_i^0(t)\mathbf{a}_i^1(x;y)$.

A particular case of interest concerns *affine parametric* expansions for the convection field

$$\mathbf{a}(t,x;y) = \sum_{j=1}^{d_y} y_j \mathbf{a}_j(t,x), \quad \mathcal{Y} = [-1,1]^{d_y}, \tag{3.6}$$

i.e., $\mathbf{a} : \mathbb{R}^{m+1+d_y} \to \mathbb{R}^m$. Such representations arise, for instance, from Karhunen–Loéve expansions of random convection fields in which case the $\mathbf{a}_j$ have some decay properties. Notice that the second relation in (3.2) now reads

$$\|\mathbf{a}\|_{L_\infty(\Omega)} = \sup_{(t,x)\in\Omega} \sum_{j=1}^{d_y} |\mathbf{a}_j(t,x)| \le A. \tag{3.7}$$

More specifically, we choose the following representation format that allows us later to explore several possible regimes

$$\begin{aligned} \mathbf{a}_j(t,w) &= \omega_j \mathbf{a}_j^\circ(t,w), \quad \|\mathbf{a}_j^\circ\|_{L_\infty([0,\widehat{T}]\times\mathbb{R}^m)} \le A^\circ, \\ \Lambda &:= \max_{\substack{j=1,\dots,d_y \\ t\in[0,\widehat{T}]}} |\mathbf{a}_j^\circ(t,\cdot)|_{\mathrm{Lip}_1(\mathbb{R}^m)}, \end{aligned} \tag{3.8}$$

where

$$\underline{\omega} = (\omega_1,\dots,\omega_{d_y}) \in \mathbb{R}_+^{d_y}, \quad |\underline{\omega}|_1 := \sum_{j=1}^{d_y} \omega_j, \quad \text{so that} \quad A = |\underline{\omega}|_1 A^\circ.$$

Then, one has for all $(x,y),(x',y') \in \mathbb{R}^m \times \mathcal{Y}$

$$|\mathbf{a}(t,x;y) - \mathbf{a}(t,x';y')| \le \sum_{j=1}^{d_y} |y_j - y_j'||\mathbf{a}_j(t,x)| + |y_i'||\mathbf{a}_j(t,x) - \mathbf{a}_j(t,x')|$$

$$\leq A|y-y'| + \Lambda|\underline{\omega}|_1|x-x'|,$$

and therefore

$$\sup_{t\in[0,\widehat{T}]} |\mathbf{a}(t,\cdot;\cdot)|_{\mathrm{Lip}_1(\mathbb{R}^m\times\mathcal{Y})} \leq A + \Lambda|\underline{\omega}|_1 =: L \tag{3.9}$$

is a valid Lipschitz constant permitted in (3.2). Note that we *do not* require here the validity of (3.3).

Note also that (3.7) is possible even when $\omega_j = 1$, i.e., $|\underline{\omega}|_1 = d_y$ in which case A and L are proportional to d_y. The case $|\underline{\omega}|_1 = 1$ ensures a dimension-independent boundedness and regularity of the convection field.

However, in either of the two "extreme" regimes (R1): $|\underline{\omega}|_1 = 1$, (R2) : $|\underline{\omega}|_1 = d_y$, the convection field $\mathbf{a}$ is m-dimension sparse. More specifically, one has:

Remark 3.2 Assume that the convection field $\mathbf{a}$ is of the form (3.6), satisfying (3.7) and (3.8). Then $\mathbf{a}$ has an m-dimension sparse compositional representation of depth two

$$\mathbf{a}(t,\cdot;\cdot) \in \mathfrak{C}_{N_\mathbf{a},m}, \quad \mathrm{depth}(\mathbf{a}) = 2, \quad \mathrm{N_a} := \mathfrak{N}(\mathbf{a}) = \mathrm{d_y}(1+\mathrm{m}^2)+1, \tag{3.10}$$

and

$$|\!|\!|\mathbf{a}(t,\cdot,\cdot)|\!|\!|_{N,m} \leq A + \Lambda|\underline{\omega}|_1, \quad N \geq N_\mathbf{a}. \tag{3.11}$$

Hence $\mathbf{a}$ belongs to $L_\infty(0,\widehat{T};\mathcal{A}^{\gamma,m})$ for *every* growth function γ.

To see this, note that

$$\mathbf{a}(t,x;y) = (g^2\circ g^1)(t,x;y),$$

where, in view of (2.2), for $S_A := \{(r^1,\dots,r^{d_y}) \in \mathbb{R}^{md_y} : \sum_{j=1}^{d_y}|r^j| \leq A\}$,

$$\left.\begin{aligned} g^1 &: (t,x,y) \mapsto (y, \mathbf{a}_1(t,x),\dots,\mathbf{a}_{d_y}(t;x)) \in \mathbb{R}^{d_y(1+m)},\\ g^2 &: (y,r^1,\dots,r^{d_y}) \in \mathcal{Y}\times S_A \mapsto \textstyle\sum_{j=1}^{d_y} y_j r^j, \end{aligned}\right\} \quad \mathfrak{N}(\mathbf{a}) = d_y(1+m^2)+1.$$

This shows (3.10). Moreover, we infer from (3.9) and (3.7) that

$$|g^1|_{\mathrm{Lip}_1} \leq \max\{1,\Lambda\}, \quad , |g^2|_{\mathrm{Lip}_1(\mathcal{Y}\times S_A)} \leq \big(A+\Lambda|\underline{\omega}|_1\big)\max\big\{|r-r'|,|y-y'|\big\}.$$

Since by assumption $|g^2\circ g^1|_{\mathrm{Lip}_1} = |\mathbf{a}|_{\mathrm{Lip}_1} \leq \big(A+\Lambda|\underline{\omega}|_1\big) = L$ we see that uniformly in t, as a function of x, y, one has $\mathbf{a} \in \mathfrak{C}_{N_\mathbf{a},m}$ where $N_\mathbf{a} := \mathfrak{N}(\mathbf{a}) = d_y(1+m^2)+1$. This confirms (3.11). □

Remark 3.3 To reduce technicalities when tracking the dependence of constants on problem parameters we assume from now on that

$$1 \leq L_t, A \leq L, \tag{3.12}$$

because a large L will be seen to have the most adverse effect. Finally, recall that, by definition

$$A, L \leq \|\mathbf{a}\|_{L_\infty(0,\widehat{T};\mathcal{A}^{\gamma,s})} =: \|\mathbf{a}\|. \tag{3.13}$$

where this latter notational abbreviation will be used whenever reference to γ, s is clear from the context.

3.3 Characteristics

The field of characterics, given by the family of ODEs

$$\dot{z}(t) = \mathbf{a}(t, z(t); y), \quad z(0) = x, \tag{3.14}$$

plays a pivotal role in what follows. Note that the characteristics have a natural semi-group property, namely that they can be obtained by composing individual characteristic segments. More precisely, suppressing the dependence on y for a moment, we consider the solution of the more general initial value problem

$$\dot{z}(t, \tau; \bar{z}) = \mathbf{a}(t, z(t)), \quad z(\tau) = \bar{z}. \tag{3.15}$$

Later concatenations of characteristic segments necessitates including a specific initial time τ in the notation. If $\tau = 0$ and there is no risk of confusion we often abbreviate $z(t, 0; \bar{z}) = z(t; \bar{z})$. Thus, one has for any τ

$$z(t, x) = z(t, \tau; z(\tau; x)) =: (z(\cdot; \tau; \cdot) \circ z(\cdot; 0; x))(t). \tag{3.16}$$

In slight abuse of terminology we refer to this as *composing* characteristic segments.

There is a second angle regarding compositional approximations to characteristics, namely that (3.14) is equivalent to the *fixed-point* relation

$$z(t, \tau; x; y) = x + \int_\tau^t \mathbf{a}(s, z(s; \tau; x; y); y) ds. \tag{3.17}$$

Both, the semi-group property and the fixed-point relation will be combined to construct compositional approximations to the characteristics.

Under the above assumptions characteristics don't cross, i.e., the value of the solution to (3.6) can be determined by tracing back along characteristics. In fact, in view of (3.14), one has for the solution u of (3.1) (suppressing again the dependence on y for a moment) $\frac{d}{dt} u(t, z(t)) = f(t, z(t))$. Hence, recalling that $u(0, z(0, x; y); y) = u_0(x; y)$,

$$u(t, z(t, x; y), y) = u_0(x; y) + \int_0^t f(s, z(s, x; y)) ds, \tag{3.18}$$

or equivalently, using (3.16) and noting that when $x = z(t, 0; \bar{x})$ one has $z(s, 0; \bar{x}) = z(-(t-s), t; x) = z(s-t, t; x)$, (3.18) takes the form

$$u(t, x, y) = u_0(z(-t, t; x, y), y) - \int_0^t f(s, z(s-t, t, x; y))ds. \tag{3.19}$$

In summary, if the charateristics have "good (pointwise) compositional approximability", for $f = 0$, the solution results from one additional compostion.

The central objective in what follows is to construct finitely parametrized surrogates $\mathcal{N}(t, z, y)$ for the map

$$(t, z, y) \in \Omega \mapsto u(t, z, y),$$

that are determined by possibly few degrees of freedom. The general flavor of the following results is: membership of the problem data (convection field, initial conditions, right hand side) to an approximation class (see Sect. 2.5) implies membership of characteristics and solution to a certain approximation class.

4 Main Results

In view of the representations (3.18), (3.19) of solutions to the parametric transport equation (3.1), the first group of results in Sect. 4.1 is devoted to establishing dimension-sparse compositional approximability of parametric characteristic fields, provided that the convection field permits dimension sparse compositional approximations. It is perhaps worth stressing that these results are *not* obtained by discretizing the characteristic ODE (3.14) which would yield much weaker results, see related comments in Sect. 6.1 below.

Section 4.2 is then devoted to the main "regularity theorem" quantifying how dimension-sparse approximability of the problem data implies dimension sparse approximability of the solution.

The basic architecture of the proofs in Sect. 4.1 is to first quantify, dimension sparse compositional approximability of characteristics by combining the semi-group property (3.16) with the fixed point property (3.17). Note that the lengths of the underlying characteristic segments—macro time steps, so to speak—depends only on problem parameters L, A, $\|\mathbf{a}\|$, but not on the target accuracy which determines the number of fixed point iterations.

In favor of an easier interpretability we focus on the exemplary types of growth functions $\gamma \sim$ (alg) and $\gamma \sim$ (exp), defined in (2.24).

Since the spatial dimension m is fixed and at most three we do not always mark the dependence of estimates on m. The general message is that compositional approximability of the problem data (convection field, initial condition, right hand side) is inherited by the characteristic fields and solutions. These results can be viewed as "regularity theorems" in a broad sense. They culminate in quantifying approximability of *solution operators*.

The proofs of the following results can all be found in Sect. 6.

4.1 Compositional Approximability of Characteristic Fields

The point of the first result is to establish "regularity" of characteristics from "regularity" of the convection field, both times in terms of s-dimension sparse compositional approximability.

Theorem 4.1 *Let $\widehat{I} := [0, \widehat{T}]$ and assume that the convection field $\mathbf{a}$ satisfies (3.4) as well as (3.5) or (3.3) for some growth function γ of either type in (2.24). Abbreviating as before $\|\mathbf{a}\| := \|\mathbf{a}\|_{L_\infty(\widehat{I};\mathcal{A}^{\gamma,\mathrm{s}})}$, one has*

$$z \in \mathrm{Lip}_1(\Omega) \cap L_\infty(\widehat{I}; \mathcal{A}^{\widetilde{\gamma},\mathrm{s}}), \quad \|z\|_{L_\infty(\widehat{I};\mathcal{A}^{\widetilde{\gamma},\mathrm{s}})} \lesssim e^{\|\mathbf{a}\|\widehat{T}}.$$

where

$$\widetilde{\gamma}(r) := \begin{cases} \left(\frac{rC_a^{1/\alpha}}{A\widehat{T}}\right)^{\frac{\alpha}{1+\alpha}} \left(\log_2\left(\frac{rC_a^{1/\alpha}}{A\widehat{T}}\right)\right)^{-\frac{\alpha}{1+\alpha}}, & \gamma \sim (alg), \\ \frac{\alpha r}{A\widehat{T}}\left(\log_2\left(\frac{\alpha r}{A\widehat{T}}\right)\right)^{-2}, & \gamma \sim (exp). \end{cases} \tag{4.1}$$

In particular, the parameter dependent characteristic field satisfies

$$\inf_{\substack{\mathcal{C}\in\mathfrak{C}_{N,\mathrm{s}} \\ \|\mathcal{C}\|_{N,\mathrm{s}}\le e^{\widehat{T}\|\mathbf{a}\|}}} \|z - \mathcal{C}\|_{L_\infty(\Omega)} \lesssim e^{\widehat{T}\|\mathbf{a}\|}\widetilde{\gamma}(N)^{-1}, \quad N \in \mathbb{N}. \tag{4.2}$$

The tamed compositional approximations in (4.2) are not yet characterized by a finite number of degrees of freedom which is done in a next step similar to Theorem 2.16.

In what follows we adopt a generous understanding of deep neural networks regarding the dependence on the time variable t. We allow in essence layers that are piecewise affine in t and hence still enjoy the basic properties of DNNs regarding evaluation and back-propagation.

Theorem 4.2 *Under the same assumptions on the convection field $\mathbf{a}$ and growth functions γ according to (2.24) there exists for each $\varepsilon > 0$ a deep neural network (DNN) $\mathcal{N}_\varepsilon$ such that*

$$\|z - \mathcal{N}_\varepsilon\|_{L_\infty(\Omega;\mathbb{R}^m)} \le \varepsilon,$$

and

$$\#\mathcal{N}_\varepsilon \lesssim A\widehat{T}2^{\mathrm{s}}\|\mathbf{a}\|^{2\mathrm{s}} \begin{cases} C_a^{-\frac{1}{\alpha}}\left(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\right)^{\frac{(1+\mathrm{s})(1+\alpha)}{\alpha}} \left|\log_2\left(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\right)\right|^2, & \gamma \sim (alg), \\ \alpha^{-(1+\mathrm{s})}\left(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\right)^{1+\mathrm{s}} \left|\log_2\left(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\right)\right|^{3+\mathrm{s}}, & \gamma \sim (exp). \end{cases} \tag{4.3}$$

It is instructive to reformulate Theorem 4.2 in terms of *convergence rates*.

Corollary 4.3 *For each $N \in \mathbb{N}$, there exists a DNN $\mathcal{N}_N$ of complexity $\#\mathcal{N}_N \leq N$, such that*

$$\|z - \mathcal{N}_N\|_{L_\infty(\Omega;\mathbb{R}^m)} \lesssim e^{\widehat{T}\|\mathbf{a}\|}\widetilde{\gamma}(N)^{-1},$$

where

$$\widetilde{\gamma}(r) \asymp \begin{cases} B\, r^{\frac{\alpha}{(1+\alpha)(1+\mathrm{s})}} \left|\log_2 r\right|^{-\frac{2\alpha}{(1+\alpha)(1+\mathrm{s})}}, & \text{when } \gamma(r) \sim r^\alpha, \\ C\alpha\, r^{\frac{1}{1+\mathrm{s}}} \left|\log_2 r\right|^{-\frac{3+\mathrm{s}}{1+\mathrm{s}}}, & \text{when } \gamma(r) \sim e^{\alpha r}, \end{cases} \tag{4.4}$$

with

$$B = C_a^{\frac{1}{(1+\mathrm{s})(1+\alpha)}} (A\widehat{T}2^{\mathrm{s}}\|\mathbf{a}\|^{2\mathrm{s}})^{-\frac{\alpha}{(1+\mathrm{s})(1+\alpha)}}, \quad C = \left(A\widehat{T}2^{\mathrm{s}}\|\mathbf{a}\|^{2\mathrm{s}}\right)^{-\frac{1}{1+\mathrm{s}}}.$$

Remark 4.4 In both theorems the exponential case can be seen as a formal "limit $\alpha \to \infty$" of algebraic rates. For $\mathrm{s} = m + d_y$ the obtained rate would reflect the full CoD. The above bounds do not show an explicit dependence on the parametric dimension d_y. In particular, the CoD does not show in the rate, i.e., through a d_y-dependent power of ε^{-1}. However, a dependence on d_y could be hidden in the quantity $\|\mathbf{a}\|$, defined in (3.13). A related example is detailed in Corollary 4.6 below.

Remark 4.5 It seems that one cannot expect in general a uniform bound on the composition norms $|||\mathcal{N}_\varepsilon|||_{\#\mathcal{N}_\varepsilon,\mathrm{s}}$, see Theorem 2.16 and the comments preceding Lemma 6.7 in Sect. 6.4, unless the compositional approximations of $a(t,\cdot;\cdot)$ have uniformly bounded depth. For the growth-types (alg) and (exp) in (2.24), the following holds

$$|\mathcal{N}_\varepsilon|_{\mathrm{Lip}_1([0,\widehat{T}];\mathbb{R}^m\times\mathcal{Y})} \lesssim \max\{1, \|\mathbf{a}\|\}, \quad |||\mathcal{N}_\varepsilon|||_{\#\mathcal{N}_\varepsilon,\mathrm{s}} \lesssim e^{L_\varepsilon \widehat{T}}, \quad \varepsilon > 0,$$

where

$$L_\varepsilon \lesssim \begin{cases} (c_3(1+A)\|\mathbf{a}\|)^{(2/C_a)^{\frac{1}{\alpha}}\varepsilon^{-1/\alpha} e^{\|\mathbf{a}\|\widehat{T}/\alpha}}, & \text{in case } \gamma \sim \text{(alg)}, \\ (c_3(1+A)\|\mathbf{a}\|)^{\frac{1}{\alpha}\left(|\ln\varepsilon|+\|\mathbf{a}\|\widehat{T}\right)}, & \text{in case } \gamma \sim \text{(exp)}. \end{cases}$$

Here c_3 is the constant from (2.19). Thus, Lipschitz continuity with respect to x, y degrades when ε decreases, the less though, the stronger the growth order of γ.

Recall that convection fields with affine parameter dependence arise, for instance, through (truncated) Karhunen–Loeve expansions of random convection fields. Remark 3.2 says that $\mathbf{a}$ belongs then to $C(\widehat{I};\mathcal{A}^{\gamma,m})$ for any growth function γ. A related first result follows from Theorem 4.2 and Corollary 4.3 by a judicious choice of γ.

Corollary 4.6 *Assume that $\mathbf{a}$ is of the form (3.6) and satisfies (3.7) and (3.8). Then,*

$$\|\mathbf{a}\| = \|\mathbf{a}\|_{L_\infty(\widehat{I};\mathcal{A}^{\gamma,m})} \leq 2A + \Lambda|\underline{\omega}|_1, \tag{4.5}$$

and the characteristic field belongs to $\mathrm{Lip}_1(\Omega)\cap C(\widehat{I};\mathcal{A}^{\widetilde{\gamma},m})$ *where*

$$\widetilde{\gamma}(r)\asymp \frac{r}{d_y A\widehat{T}}\Big|\log_2\frac{r}{d_y A\widehat{T}}\Big|^{-2},\quad \|z\|_{L_\infty([0,\widehat{T}];\mathcal{A}^{\widetilde{\gamma},m})}\le e^{(2A+\Lambda|\underline{\omega}|_1)\widehat{T}}.$$

Moreovery, for each $N\in\mathbb{N}$ *there exists a network* $\mathcal{N}_N$ *such that*

$$\|z-\mathcal{N}_N\|_{L_\infty(\Omega\times\mathcal{Y})}\lesssim d_y F e^{(2A+\Lambda|\underline{\omega}|_1)\widehat{T}}N^{-\frac{1}{m+1}}\big|\log_2 N\big|^{\frac{3+m}{1+m}},\tag{4.6}$$

where $F=(A\widehat{T}2^m(2A+\Lambda|\underline{\omega}|_1)^{2m})^{\frac{1}{m+1}}$.

While the *rates* do not suffer from the CoD, to gain traction, N has to exceed d_y^{m+1}. Although this delay effect is only algebraic in d_y, this dependence is not optimal since the choice of any growth function for $\mathbf{a}$ does not fully exploit the special structure (3.6), see the proof in Sect. 6.5. A more direct reasoning yields the following better results with regard to the stability of the networks, the scaling in $\widehat{T}$, and the dependence on d_y.

Theorem 4.7 *Assume that* $\mathbf{a}$ *is of the form* (3.6) *and satisfies* (3.7) *and* (3.8). *Recall from* (3.9) *that*

$$L:=A+\Lambda|\underline{\omega}|_1.\tag{4.7}$$

Then, for any $\varepsilon>0$ *there exists a DNN* $\mathcal{N}_\varepsilon$ *such that*

$$\|z-\mathcal{N}_\varepsilon\|_{L_\infty(\Omega;\mathbb{R}^m)}\le\varepsilon,\quad \#\mathcal{N}_\varepsilon\lesssim d_y m^2 A\widehat{T}\Big(\frac{e^{L\widehat{T}}}{\varepsilon}\Big)^{m+1}\Big|\log_2\frac{e^{L\widehat{T}}}{\varepsilon}\Big|^2.\tag{4.8}$$

Moreover, there exists a DNN $\mathcal{N}_N$ *with complexity* $\#\mathcal{N}_N\le N$ *such that*

$$\begin{aligned}\|z-\mathcal{N}_N\|_{L_\infty(\Omega;\mathbb{R}^m)}&\lesssim e^{L\widehat{T}}\widetilde{\gamma}(N)^{-1},\ N\in\mathbb{N},\\ \widetilde{\gamma}(r)&=C\Big(\frac{r}{d_y}\Big)^{\frac{1}{m+1}}\Big|\log_2\frac{r}{d_y}\Big|^{-\frac{2}{m+1}},\end{aligned}\tag{4.9}$$

where $C=(A\widehat{T}m^2)^{-\frac{1}{m+1}}$. *The networks belong to* $\mathrm{Lip}_1([0,\widehat{T}];C(\Omega;\mathbb{R}^m))$ *and are stable with* $|||\mathcal{N}_N|||_{N,m}\lesssim e^{\hat{L}\widehat{T}}$ *where* $\hat{L}\le A+\widehat{T}^{-1}+c_3(1+A^\circ)\Lambda|\underline{\omega}|_1$ *whenever* $\varepsilon\le 1$.

Remark 4.8 If, on the other hand, we consider regime (R2) $|\underline{\omega}|_1=d_y$ the Lipschitz constant $\|\mathbf{a}\|\ge L$ scales like d_y so that the constant $e^{\widehat{T}\|\mathbf{a}\|}$ depends exponentially on d_y (see (3.9)). Hence, the CoD still strikes through an exponential delay in gaining accuracy.

4.2 Approximability of the Solution Operator

We discuss next approximability of the parameter dependent solutions themselves. The following result quantifies approximability of the *data-to-solution operator* $\mathcal{S}$: $(\mathbf{a}, u_0, f) \to u$ and shows under which conditions on the data approximating $\mathcal{S}$ avoids the CoD.

Theorem 4.9 *Under the same hypotheses on the convection field* $\mathbf{a}$ *as in Theorem 4.7 assume that the data* u_0, f *satisfy*

$$u_0 \in \mathcal{A}^{\gamma,m}, \quad f \in L_\infty(\widehat{I}; \mathcal{A}^{\gamma,m}) \cap \mathrm{Lip}_1(\widehat{I}; C(\mathbb{R}^m \times \mathcal{Y})), \quad \gamma(r) \sim r^\alpha, \tag{4.10}$$

and let

$$\beta := \max\left\{1, \frac{m+1}{\alpha}\right\}. \tag{4.11}$$

Then, for any $\varepsilon > 0$ *there exsits a DNN* $\mathcal{N}_{u,\varepsilon}$ *such that for the exact solution* u *of the transport equation (3.1)*

$$\|u - \mathcal{N}_{u,\varepsilon}\|_{L_\infty(\Omega)} \le \varepsilon, \quad \#\mathcal{N}_{u,\varepsilon} \lesssim B d_y \Big(\frac{e^{\widehat{T}L}}{\varepsilon}\Big)^{(m+1+\beta)} \Big|\log_2 \frac{e^{L\widehat{T}}}{\varepsilon}\Big|^2, \tag{4.12}$$

where B *depends on* $m, L, \alpha, \max\{1, \|u_0\|, \|f\|_0\}$ *with* $\|u_0\| := \|u_0\|_{\mathcal{A}^{\gamma,m}}$, $\|f\| := \|f\|_{L_\infty(\widehat{I};\mathcal{A}^{\gamma,m})}$.

Moreover, for $N \in \mathbb{N}$ *there exists a stable DNN* $\mathcal{N}_N$ *with* $\#\mathcal{N}_N \le N$ *such that*

$$\|u - \mathcal{N}_N\|_{L_\infty(\Omega)} \lesssim e^{L\widehat{T}} \underbrace{(d_y B)^{\frac{1}{m+1+\beta}} N^{-\frac{1}{m+1+\beta}} \Big|\log_2 \frac{N}{B}\Big|^{\frac{2}{m+1+\beta}}}_{:=\hat{\gamma}(N)^{-1}}, \tag{4.13}$$

with $|||\mathcal{N}_N|||_{N,m} \lesssim \max\{1, \|u_0\|, \|f\|\} e^{L\widehat{T}}$. *Thus, for* $\hat{\gamma}$ *defined in (4.13), we have* $u \in C(\widehat{I}; \mathcal{A}^{\hat{\gamma},m}) \cap \mathrm{Lip}_1(\widehat{I}; C(\mathbb{R}^m \times \mathcal{Y}))$. *In other words, dimension sparse approximability of the problem data implies a certain quantified dimension sparse approximability of the solution.*

Remark 4.10 Although irrespective of the CoD the rate (4.13) becomes arbitrarily bad (β becomes arbitrarily large) when the algebraic order α gets small below the space-time dimension $m+1$. The smallest value $\beta = 1$ for $\alpha \ge m+1$, as opposed to a value tending to zero when α grows as in (4.4), is due to the additional time-integration on the source field f. If one replaces the algebraic order $\gamma(r) \sim r^\alpha$ in (4.10) by an exponetial growth order $\gamma(r) \sim e^{\alpha r}$ one can show that

$$\|u - \mathcal{N}_{u,\varepsilon}\|_{L_\infty(\Omega)} \le \varepsilon, \quad \#\mathcal{N}_{u,\varepsilon} \lesssim B d_y \Big(\frac{e^{\widehat{T}L}}{\varepsilon}\Big)^{(m+2)} \Big|\log_2 \frac{e^{L\widehat{T}}}{\varepsilon}\Big|^2,$$

(corresponding to $\beta = 1$) with the same dependencies of B on problem parameters. Finally, we could have replaced $f \in \mathrm{Lip}_1(\widehat{I}; C(\mathbb{R}^m \times \mathcal{Y}))$ by an assumption like (3.5).

5 Comments and Outlook

The common trait of the above results is a uniform approximation rate for the characteristic field close to $\varepsilon^{-\frac{1}{m+1}}$ (the closer the stronger the approximability order of the convection field). This is the rate one can expect for a ball in $\mathrm{Lip}_1(\widehat{I} \times D; \mathbb{R}^m)$, i.e., functions of $m+1$ variables. For the solutions themselves it seems that one cannot quite benefit from increasing algebraic growth orders beyond $\alpha = m+1$. Since again the range of possible solutions u is dense in a Lipschitz ball of $\mathrm{Lip}_1(\widehat{I} \times D)$ the obtained rate in Theorem 4.9 seems to be close to optimal. Moreover, whenever the problem data have some compositional dimension-sparsity, in all scenarios the constructed approximations avoid the CoD. In general, emphasis has been on weak dependence on d_y *not* on high order rates which, from the perspective of practical realization would be a questionable goal anyway, [18] (see also further comments later below).

We conclude with indicating some ramifications of the preceding findings whose detailed treatment is postponed to forthcoming work. Let $\mathcal{M}(\mathbf{a}, \mathcal{Y})$ denote the set of characteristic fields $z(\cdot, \cdot; y)$ obtained when y traverses $\mathcal{Y}$ for a fixed given convection field $\mathbf{a}$, while the solution manifold $\mathcal{M}(\mathbf{a}, u_0, f, \mathcal{Y})$ is comprised of all solutions to (3.1) for fixed data $\mathbf{a}$, u_0, f. To capture stability with respect to those data as well, let $\mathfrak{A}$ denote the set of all convection fields with fixed bounds for L, L_t, A, $\|\mathbf{a}\|_{L_\infty(\widehat{I};\mathcal{A}^{\gamma,s})}$. Likewise let $\mathfrak{F}$ denote the set of all (u_0, f) with $\|u_0\|_{\mathcal{A}^{\gamma,m}}, \|f\|_{L_\infty(\widehat{I};\mathcal{A}^{\gamma,m})} \le M$. Obviously, $\mathfrak{A}$, $\mathfrak{F}$ are compact in $C(\Omega)$. For the Lipschitz-regularizer $\mathcal{R}$ from (2.14) the preceding results say that all sets

$$\begin{gathered}\mathcal{M}(\mathbf{a}, \mathcal{Y}),\ \ \mathcal{M}(\mathbf{a}, u_0, f, \mathcal{Y}),\quad \mathcal{M}(\mathfrak{A}, \mathcal{Y}) := \bigcup_{\mathbf{a}\in\mathfrak{A}} \mathcal{M}(\mathbf{a}, \mathcal{Y}),\\ \mathcal{M}(\mathfrak{A}\times\mathfrak{F}, \mathcal{Y}) := \bigcup_{(\mathbf{a},u_0,f)\in\mathfrak{A}\times\mathfrak{F}} \mathcal{M}(\mathbf{a}, u_0, f, \mathcal{Y}),\end{gathered} \tag{5.1}$$

are contained in bounded balls of spaces of the type $L_\infty(\widehat{I}; \mathcal{A}^{\widetilde{\gamma},s}) \cap \mathrm{Lip}_1(\Omega)$ for some growth function $\widetilde{\gamma}$.

As indicated earlier in the introduction, a common way of characterizing the complexity of these collections is to determine their *metric entropy* or suitable versions of nonlinear *widths*, among those so-called (nonlinear) *manifold widths*, introduced in [11], see (1.3). Denoting by $\theta_{N,\cdot} \in \mathbb{R}^N$ the collection of weights defining the respective DNN approximations $\mathcal{N}_{N,z}, \mathcal{N}_{N,u}$ in Theorems 4.2, 4.7, 4.9, respectively, the functions

$$\begin{aligned}D_N(t, x; \theta_N(\mathbf{a}; y)) &:= \mathcal{N}_{N,z}(t, x; y; \theta_{N,z}),\\ D_N(t, x; \theta_N(\mathbf{a}, y, u_0, f)) &:= \mathcal{N}_{N,u}(t, x; y; \theta_{N,u}(u_0; f)),\end{aligned}$$

are valid candidates for encoder-decoder pairs $D_N \circ E_N$, where $E_N(\mathbf{a}, y) = \theta_N(\mathbf{a}; y) \in \mathbb{R}^N$, $E_N(\mathbf{a}, y, u_0, f) = \theta_N(\mathbf{a}, y, u_0, f)) \in \mathbb{R}^N$, are the mappings that take $z(\cdot, \cdot; y)$, respectively $u(\cdot, \cdot; y)$ into $\theta_N(\mathbf{a}; y)$, $\theta_N(\mathbf{a}, y, u_0, f)$. Confining the discussion to $\mathbf{a}$ according to (3.6), for $\mathcal{K} \in \{\mathcal{M}(\mathbf{a}, \mathcal{Y}), \mathcal{M}(\mathfrak{A}, \mathcal{Y})\}$ the continuity of E_N, D_N can be established based on the presented results. In fact, continuity in y follows from the constructive proofs which is all that is needed for fixed $\mathbf{a} \in \mathfrak{A}$. As a next step, continuity in $\mathbf{a} \in \mathfrak{A}$ follows from the continuity of the construction of the

implanted Lipschitz stable networks from Proposition 2.14, as can be seen by inspecting the proof in Appendix A. To extend these arguments to the remaining sets in (5.1), one yet has to establish the existence of continuous metric (or near metric) selections on the level of dimension sparse compositional approximations prior to implanting Lipschitz stable DNNs. In particular, this would yield bounds for the manifold widths of compact sets of the type (2.11).

Knowing the manifold-widths does not allow one to infer directly on the entropy numbers of the sets in (5.1) (and hence on the number of bits needed to encode the centers of respective ε-covers). For a strenghtened version of manifold widths, so called *stable widths*, introduced in [5], a version of Carl's inequality is known which asserts that an algebraic order of stable widths implies the corresponding algebraic order of the entropy numbers. These stable widths require both factors E_N, D_N to be Lipschitz continuous. For fixed $\mathbf{a}$, u_0, f, the above findings assert (uniform) Lipschitz continuity of the compositions $D_N \circ E_N$ (for $\mathbf{a}$ of the type (3.6)). It is known that DNNs are Lipschitz continuous with respect to the weights under size constraints on the weights, see e.g. [30]. In general corresponding Lipschitz constants are expected to be very large which impedes an inference from approximation rates to entropy numbers. This gives rise to the notion of Lipschitz widths studied in [30]. There, among other things, bounds on entropy numbers are derived from DNN approximation rates which are (necessarily) somewhat weaker than those in Carl's inequality, see [30, § 6.2]. Since they are derived under specific architecture constraints (either widths or depths stay bounded) they do not apply directly to the scenarios discussed here. Specifying (and perhaps refining) such results to the current situation would be interesting as they may shed light on how the entropy numbers of the solution manifolds in (5.1) relate to those of the accommodating balls of type (2.11).

Finally, dimension-sparse DNN-approximability implies compositional approximability with the same dimension sparsity. By Theorem 2.16, the converse is also true. However, the respective growth orders γ don't necessarily match. It would be interesting to further narrow this gap.

In a different direction, in principle, the framework allows us to treat even less regular data leading to solutions that are no longer Lipschitz continuous. This may require weaker regularizations than (2.14) or refined notions of approximation classes that allow gradually increasing Lipschitz constants in compositional approximations, as indicated in Remarks 2.7, 2.6. Remark 2.17 already indicates a wider scope of applications. For instance, it would be interesting to apply the above concepts to nonlinear conservation laws by exploiting their equivalent kinetic formulations as linear parametric transport equations, see [25]. An obvious obstacle here is that the right hand sides are measure-valued. However, solutions do satisfy linear transport equations with zero right hand side on regions separated by shocks. Alternatively, one may consider constructing compositional approximations generated through the fixed-point iterations considered in [36]. Splitting methods for more involved kinetic models may serve as another starting point for generating compositional approximations. Finally, the above concepts apply as well to high-dimensional transport equations and solution manifolds induced by source terms and initial conditions. Aside from their role in Fokker-Planck equations, the correspondence bewteen nonlinear high-dimensional dynamical systems and linear transport PDEs opens another interesting perspective.

Finally, one may consider the case of smooth data for which one could expect better rates. However, in the end one may have to resort to training concepts, typically based on point samples to determine DNN approximations, perhaps in combination with pre-structured architectures suggested by the constructive proofs. It has been shown, however, in [18] that there is no hope then to realize higher convergence orders.

6 Proofs for Sect. 4

6.1 Road Map

The proofs are organized in two groups of intermediate results:

(A) technical prerequisits (Sect. 6.2);

(B) the actual proofs of the results in Sect. 4 (Sects. 6.3–6.6).

We collect in (A) several tools that are used *repeatedly* in (B). Here is a brief overview:

Section 6.2.1 provides a mechanism how to reformulate ε-accurate error bounds in terms of convergence rates $O(\widetilde{\gamma}(N)^{-1})$ with a corresponding growth function $\widetilde{\gamma}$.

Section 6.2.2: The analysis of compositional approximability of the characteristics is based on two constituents, namely the semi-group property of solutions to the characteristic ODE (3.14) and on the fact that the characteristics solve a fixed point equation (3.17). In this section we construct an approximation to characteristics through an appropriate number of fixed point iterations on judiciously chosen characteristic segments. We stress that the length of these segments depends only on properties of the convection field, *not* on the target accuracy. Hence, the resulting compositional approximation is not obtained by discretizing the ODE in a straightforward way.

The resulting composition of operators needs to be approximated by compositions of pointwise mappings which is done, step by step, in the subsequent subsections.

Section 6.2.3: Remark 2.13 will later be used to control the error incurred by approximating the operator composition. This requires bounds for the Lipschitz constants for iterates of the fixed point operator on each characteristic segment which are derived in this section.

Section 6.2.4: Integral operators arise in the fixed point relation for characteristics but also in the solution representation (3.18) or (3.19). This section presents approximation of integral operators by pointwise compositions, based on elementary quadrature. This is shown to generate approximations of the fixed point operator by compositions of pointwise mappings.

Section 6.2.5: These approximations still involve the exact convection field. In a further approximation step the convection field is approximated by a dimension sparse compositional approximation. This is used in Sect. 6.3 as well as later as the starting point for implanting DNNs for the proof of Theorem 4.2 in Sect. 2.3, always using Remark 2.13.

Section 6.2.6: This section provides Lipschitz bounds for the final compositional approximation to the characteristic field, which are later needed in the proof of Theorem 4.9.

With the above prerequisits at hand, the proofs of the results in Sect. 4 are presented (in their order) in Sects. 6.3–6.6. Specifically, the proof of Theorem 4.2 builds on the regularity theorem 4.1, again making use of Remark 2.13 and Lemma 2.1.

Section 6.5 is devoted to the more special convection fields with affine parameter dependence. As shown in Corollary 4.6 this can be viewed as an application of Theorem 4.2. However, in view of the importance of this case, we provide in addition the derivation of sharper results (Theorem 4.7). The techniques are similar in spirit but require separate arguments.

6.2 Some Technical Prerequisits

6.2.1 "Inverting" Growth Functions

Remark 6.1 Given $v \in \mathbb{X}$, suppose we have found for each $\varepsilon > 0$ an approximation v_ε, depending on at most N_ε degrees of freedom, that satisfies $\|v - v_{N_\varepsilon}\|_{\mathbb{X}} \leq \varepsilon$. If $N_\varepsilon \eqsim \phi(Q/\varepsilon)$ for some strictly increasing function ϕ of at most algebraic growth, then one has

$$\|v - v_N\|_{\mathbb{X}} \lesssim Q\gamma(N)^{-1}, \quad N \in \mathbb{N}, \tag{6.1}$$

where $\gamma(r)$ is any growth function satisfying

$$\gamma(\phi(s)) \eqsim s. \tag{6.2}$$

We often briefly write then $\gamma \eqsim \phi^{-1}$. This will be repeatedly used as follows: Suppose the v_N in (6.1) belong to $\mathfrak{C}_{N,\mathrm{s}}$ and $\|\!|g_N|\!\|_{N,\mathrm{s}} \leq Q$ for all $N \in \mathbb{N}$. Then $v \in \mathcal{A}^{\gamma,\mathrm{s}}$ with $\|v\|_{\mathcal{A}^{\gamma,\mathrm{s}}} \lesssim Q$. To see the last conclusion, just note that $\gamma(N)\Big\{\|v - v_N\|_{\mathbb{X}} + \gamma(N)^{-1}\|\!|v_N|\!\|_{N,\mathrm{s}}\Big\} \lesssim 2Q$.

Appropriate "near-inverses" γ will be needed for growth functions ϕ of the following form.

Lemma 6.1 *Assume that for positive b_1, b_2, ζ and real β*

$$\phi(s) = b_1 s^\zeta \big|\log_2 b_2 s\big|^\beta, \quad s \geq s_0 > 0.$$

Then

$$\phi^{-1}(r) \eqsim b_1^{-1/\zeta} \zeta^{\beta/\zeta} r^{\frac{1}{\zeta}} \big|\log_2(b_2^\zeta r/b_1)\big|^{-\frac{\beta}{\zeta}}, \quad r \geq r_0 > 0. \tag{6.3}$$

Proof Making the ansatz $\phi^{-1}(r) \eqsim F r^{\frac{1}{\zeta}} |\log_2(Qr)|^\theta$, we have

$$\begin{aligned} s = \phi^{-1}(\gamma(s)) &\eqsim F\big(b_1 s^\zeta \big|\log_2(b_2 s)\big|^\beta\big)^{\frac{1}{\zeta}} \Big|\log_2\big(Q b_1 s^\zeta \big|\log_2(b_2 s)\big|^\beta\big)\Big|^\theta \\ &= F\big(b_1 s^\zeta \big|\log_2(b_2 s)\big|^\beta\big)^{\frac{1}{\zeta}} \zeta^\theta \Big|\log_2\big((Q b_1)^{\frac{1}{\zeta}} s\big)\Big|^\theta \left|1 + \frac{\big|\log_2\big|\log_2(b_2 s)\big|^\beta\big|}{\big|\log_2(Q b_1 s^\zeta)\big|}\right|^\theta. \end{aligned}$$

Equating coefficients yields $F = b_1^{-\frac{1}{\zeta}}\zeta^{\frac{\beta}{\zeta}}$, $Q = \frac{b_2^\zeta}{b_1}, \theta = -\frac{\beta}{\zeta}$, which confirms the claim. □

Note that in the above situation the proportionality constants in (6.2) tend to one when the argument increases. For our purposes uniformly bounded proportionality constants suffice so that in later applications we can drop the constant $\zeta^{\beta/\zeta}$ in (6.3).

6.2.2 Fixed-Point Iterations and Composition of Characteristic Segments

In what follows we denote by $I := [\underline{t}, \bar{t}]$ a fixed time interval whose length $|I| := \bar{t} - \underline{t}$ depends on L. We fix the "macro-time-step" $|I|$ so that

$$|I|\|a\| = \frac{1}{2}. \tag{6.4}$$

For a given time horizon $\widehat{T}$ one then needs $K := \lceil \widehat{T}/|I| \rceil$ such steps and we assume for convenience that $K = \widehat{T}/|I|$ is already an integer. In addition we denote by

$$\Omega(I) := I \times \mathbb{R}^m \times \mathcal{Y} \subset \Omega$$

the spatio-parametric time-slab determined by I.

To find approximate compositions we recall the fixed-point relation (3.17) and consider the corresponding mapping $\Phi_{x,I} : L_\infty(\Omega(I); \mathbb{R}^m) \to L_\infty(\Omega(I); \mathbb{R}^m)$, defined by,

$$\Phi_{x,I}(t, \bar{z}; y) := x + \int_{\underline{t}}^{t} \mathbf{a}(s, \bar{z}(s), y)ds, \quad t \in I = [\underline{t}, \bar{t}]. \tag{6.5}$$

A natural strategy is to approximate the fixed point of (3.17) or (6.5) by iterates of the mapping $\Phi_{x,I}(t, \cdot; y)$. In this case the arguments $x, \bar{z}$ sometimes depend on each other. In fact, a natural initialization would be the constant-in-time function

$$\bar{z}_x(s) = x, \quad s \in I, \tag{6.6}$$

i.e., the initial value x is frozen in time throughout I. Then, we always use the notational convention

$$\Phi^k_{x,I}(t, \bar{z}; y) := \Phi_{x,I}\big(t, \Phi^{k-1}_{x,I}(\cdot, \bar{z}; y); y\big), \quad \bar{z} \in L_\infty(I; \mathbb{R}^m).$$

Condition (6.4) and $L \le \|\mathbf{a}\|$ say that $\Phi_{x,I}$ is a contraction in $\bar{z}$ since

$$\begin{aligned} |\Phi_{x,I}(t, \bar{z}; y) - \Phi_{x,I}(t, \bar{z}'; y)| &\le \int_{\underline{t}}^{t} |\mathbf{a}(s, \bar{z}(s); y) - \mathbf{a}(s, \bar{z}'(s); y)|ds \\ &\le (t - \underline{t})L\|\bar{z} - \bar{z}'\|_{L_\infty(I;\mathbb{R}^m)} \end{aligned}$$

$$\leq \frac{1}{2}\|\bar{z} - \bar{z}'\|_{L_\infty(I;\mathbb{R}^m)}. \tag{6.7}$$

Since by (3.2), $|z(t,x;y) - x| = \left|\int_{\underline{t}}^{t} \mathbf{a}(s, z(s;x);y)ds\right| \leq (t-\underline{t})A \leq A|I| \leq \frac{1}{2}$, this implies

$$\begin{aligned}
&|z(t,x;y) - \Phi^k_{x,I}(t,\bar{z}_x;y)| \\
&\quad = |\Phi_{x,I}(t, z(\cdot;x);y) - \Phi_{x,I}\big(t, \Phi^{k-1}_{x,I}(\cdot,\bar{z}_x;y)\big)| \\
&\quad \leq 2^{-k}\|z(\cdot,x;y) - \bar{z}_x(s)\|_{L_\infty(I;\mathbb{R}^m)} \leq 2^{-k}A|I| \leq 2^{-k-1},
\end{aligned}$$

where we have used $A \leq L$ (see (3.12)) and (6.4). Hence, by (6.4), it takes roughly $|\log_2 \eta|$ steps to achieve accuracy η

$$|z(t,x;y) - \Phi^{\mu}_{x,I}(t,\bar{z}_x;y)| \leq \eta, \quad (t,x,y) \in \Omega(I), \quad \mu = \mu(\eta) = \lceil \Big|\log_2(2\eta)^{-1}\Big| \rceil.$$

In view of (3.16), it is natural to concatenate next iterates $\Phi^{\mu}_{\cdot,I}$ in time for successive time intervals I. To that end, consider (for simplicity) an equally spaced partition

$$[0,\widehat{T}] = \bigcup_{k=1}^{K}[t_{k-1},t_k], \quad t_k := \frac{k\widehat{T}}{K},$$

where K is chosen in compliance with (6.4). Along with the sequence of intervals consider the vector of tolerances with corresponding sufficient iteration numbers

$$\underline{\eta}^k = (\eta_1,\ldots,\eta_k) \in \mathbb{R}^k_+, \quad \mu_k := \mu(\eta_k), \quad k = 1,\ldots,K. \tag{6.8}$$

Then define, for $j < k$, $w \in \mathbb{R}^m$

$$\begin{aligned}
&\Psi_{[k,j]}(t,w;y) := \Phi^{\mu_k}_{w_{k-1,j},I_k}(t,\bar{z}_{w_{k-1,j}};y),\ w_{k-1,j} := \Psi_{[k-1,j]}(t_{k-1},w;y),\ t \in I'_k, \\
&\Psi_{[j+1,j]}(t,w;y) := \Phi^{\mu_{j+1}}_{w,I_{j+1}}(t,\bar{z}_w;y),
\end{aligned} \tag{6.9}$$

i.e., μ_k iterates of $\Phi_{\cdot,I}$ are applied to the result of a μ_{k-1}-fold application of $\Phi_{\cdot,I}$ evaluated at the last time-junction t_{k-1}.

Specifically,

$$\Psi_{\underline{\eta}^k}(t,x;y) := \Psi_{[k,0]}(t,x;y)$$

is a natural candidate for approximating $z(\cdot,x;y)$ on I_k.

To estimate $|z - \Psi_{\underline{\eta}^k}|$ on I_k we invoke Remark 2.13. Viewing $\Psi_{\underline{\eta}^k}$ as a perturbation of the characteristic field, we need bounds for the Lipschitz constants of the exact

characteristics $z(t, w; y)$. Recall that under the above assumptions on the convection field $\mathbf{a}$, it follows from a classical Gronwall inequality that one has

$$\|z(\cdot, x; y) - z(\cdot, \bar{x}; y)\|_{L_\infty(I;\mathbb{R}^m)} \le e^{L|I|}|x - \bar{x}| \le e^{1/2}|x - \bar{x}|,$$

so that in terms of Remark 2.13 we have $L_{[k,j+1]} \le e^{L(I_{j+1}\cup\cdots\cup I_k)} = e^{L(t_k - t_j)} = e^{(k-j)/2}$. Thus, for $k \le K$, $t \in I_k$,

$$|z(t, x; y) - \Psi_{\underline{\eta}^k}(t, x; y)| \le \eta_k + \sum_{j=1}^{k-1} \eta_j e^{(k-j)/2}.$$

It remains to choose the intermediate tolerances η_j. The simplest option is to take them all equal

$$\eta_j = \eta(\varepsilon) := (e^{1/2} - 1)\varepsilon e^{-K/2}, \quad j = 1, \dots, K, \tag{6.10}$$

which yields

$$\|z(\cdot, x; y) - \Psi_{\underline{\eta}^k}(t, x; y)\|_{L_\infty(I_k;\mathbb{R}^m)} \le \varepsilon e^{(k-K)/2} \le \varepsilon, \quad k = 1, \dots, K. \tag{6.11}$$

In summary, we have

$$\|z - Z_\varepsilon\|_{L_\infty(\Omega;\mathbb{R}^m)} \le \varepsilon, \quad \text{where} \quad Z_\varepsilon(t, x; y) := \sum_{k=1}^{K} \chi_{I_k}(t)\Psi_{\underline{\eta}^k}(t, x; y). \tag{6.12}$$

6.2.3 Lipschitz Bounds

The mappings Z_ε from (6.12) are still global operators. To analyze their approximation by pointwise compositions via Remark 2.13, we need to bound the Lipschitz constants of partial compositions.

Lemma 6.2 *Under the above assumptions one has for* $k \in \mathbb{N}$, $\bar{z}, \bar{z}' \in L_\infty(I; \mathbb{R}^m)$, $y, y' \in \mathcal{Y}$

$$\begin{aligned} &|\Phi^k_{x,I}(t, \bar{z}; y) - \Phi^k_{x',I}(t, \bar{z}'; y')| \le e^{1/2}|x - x'| \\ &+ \frac{2^{-k}}{k!} \max\{\|\bar{z} - \bar{z}'\|_{L_\infty(I;\mathbb{R}^m)}, |y - y'|\}. \end{aligned} \tag{6.13}$$

In particular, one has for all $(t, x, y), (t', x', y') \in \Omega(I)$

$$\begin{aligned} &|\Phi^k_{x,I}(t, \bar{z}_x; y) - \Phi^k_{x',I}(t', \bar{z}_{x'}; y')| \le A|y - y'| \\ &+ \max\{|x - x'|, |y - y'|\}e^{1/2}, \quad k \in \mathbb{N}. \end{aligned} \tag{6.14}$$

Proof : By our assumptions (3.2) on the convection field, we conclude that for $t \in I$

$$|\Phi_{x,I}(t,\bar{z};y) - \Phi_{x,I}(t,\bar{z}';y')| \le (t-\underline{t})L \max\{\|\bar{z}-\bar{z}'\|_{L_\infty(I;\mathbb{R}^m)}, |y-y'|\},$$

so that $|\Phi_{x,I}(t,\bar{z};y)-\Phi_{x',I}(t,\bar{z}';y')| \le |x-x'|+(t-\underline{t})L\max\{\|\bar{z}-\bar{z}'\|_{L_\infty(I;\mathbb{R}^m)}, |y-y'|\}$. Hence

$$\begin{aligned}
&\Phi_{x,I}(t,\Phi_{x,I}(\cdot,\bar{z};y);y) - \Phi_{x',I}(t,\Phi_{x',I}(\cdot,\bar{z}';y');y') \\
&\quad = x + \int_{\underline{t}}^{t} \mathbf{a}\Big(s, \Big(x + \int_{\underline{t}}^{s} \mathbf{a}(s',\bar{z}(s');y)ds'\Big)\Big)ds \\
&\qquad - \Big\{x' + \int_{\underline{t}}^{t} \mathbf{a}\Big(s, \Big(x' + \int_{\underline{t}}^{s} \mathbf{a}(s',\bar{z}'(s');y')ds'\Big)\Big)ds\Big\}.
\end{aligned}$$

This yields

$$\begin{aligned}
&|\Phi_{x,I}(t,\Phi_{x,I}(\cdot,\bar{z};y);y) - \Phi_{x',I}(t,\Phi_{x'}(\cdot,\bar{z}';y');y')| \\
&\quad \le |x-x'| + \int_{\underline{t}}^{t} \Big|\mathbf{a}\Big(s, \Big(x + \int_{\underline{t}}^{s} \mathbf{a}(s',\bar{z}(s');y)ds'\Big)\Big) \\
&\qquad - \mathbf{a}\Big(s, \Big(x' + \int_{\underline{t}}^{s} \mathbf{a}(s',\bar{z}'(s');y)y')ds'\Big)\Big)\Big|ds \\
&\quad \le |x-x'| + \int_{\underline{t}}^{t} L|x-x'| + L\int_{\underline{t}}^{s} |\mathbf{a}(s',\bar{z}(s');y) - \mathbf{a}(s',\bar{z}'(s');y')|ds'ds \\
&\quad \le (1+(t-\underline{t})L)|x-x'| + L^2 \int_{\underline{t}}^{t}\int_{\underline{t}}^{s} \max\{|\bar{z}(s')-\bar{z}'(s')|, |y-y'|\}ds'ds \\
&\quad \le (1+(t-\underline{t})L)|x-x'| + \frac{((t-\underline{t})L)^2}{2} \max\big\{\|\bar{z}-\bar{z}'\|_{L_\infty(t;\mathbb{R}^d)}, |y-y'|\big\}.
\end{aligned}$$

One then easily verifies inductively that

$$\begin{aligned}
|\Phi^k_{x,I}(t,\bar{z};y) - \Phi^k_{x',I}(t,\bar{z}';y')| &\le \sum_{\nu=0}^{k-1} \frac{((t-\underline{t})L)^\nu}{\nu!}|x-x'| \\
&+ \frac{(L(t-\underline{t}))^k}{k!} \max\{\|\bar{z}-\bar{z}'\|_{L_\infty(I;\mathbb{R}^m)}, |y-y'|\},
\end{aligned} \tag{6.15}$$

which implies (6.13). Specifically, when $\bar{z}(s) = \bar{z}_x(s) = x$ for $s \in I$, (6.15)gives

$$|\Phi^k_{x,I}(t, \bar{z}_x; y) - \Phi^k_{x',I}(t, \bar{z}_{x'}; y')| \le \sum_{\nu=0}^{k} \frac{((t-\underline{t})L)^\nu}{\nu!} \max\{|x-x'|, |y-y'|\}$$
$$\le \max\{|x-x'|, |y-y'|\} e^{L|I|},$$

from which (6.14) follows for $t = t'$ since $L|I| \le 1/2$.

Since (for $t' < t$), keeping (3.2) in mind, $\left|\Phi_{x,I}(t, \bar{z}; y) - \Phi_{x,I}(t', \bar{z}, ; y)\right| \le \int_{t'}^{t} |\mathbf{a}(s, \bar{z}(s); y)| ds \le A|t-t'|$, we have

$$\left|\Phi^k_{x,I}(t, \bar{z}; y) - \Phi^k_{x,I}(t', \bar{z}; y)\right| \le \int_{t'}^{t} \left|\mathbf{a}(s, \Phi^{k-1}_{x,I}(s, \bar{z}; y); y)\right| ds$$
$$\le A|t-t'|, \quad t, t' \in I, \tag{6.16}$$

proving (6.14) and hence the assertion. □

To approximate the $\Psi_{[k,j]}$ by pointwise compositions we need the following bounds.

Corollary 6.2 *For $(t, x, y), (t', x', y') \in \Omega(I_k)$, one has*

$$|\Psi_{[k,j]}(t, w; y) - \Psi_{[k,j]}(t', w'; y')|$$
$$\le |t-t'| + e^{(k-j)/2} \max\{|w-w'|, |y-y'|\}. \tag{6.17}$$

Moreover, for Z_ε defined by (6.12), one has

$$|Z_\varepsilon(t, x, y) - Z_\varepsilon(t', x', y')| \le A|t-t'| + \max\{|x-x'|, |y-y'|\} e^{\|\mathbf{a}\|\widehat{T}}. \tag{6.18}$$

Proof Since

$$\Psi_{[k,j]}(t, w; y) - \Psi_{[k,j]}(t', w'; y') = \Phi^{\mu_k}_{w_{k-1,j}, I_k}(t, \bar{z}_{w_{k-1,j}}; y) - \Phi^{\mu_k}_{w'_{k-1,j}, I_k}(t'; \bar{z}_{w'_{k-1,j}}; y'),$$

where $w_{k-1,j} = \Psi_{[k-1,j]}(t_{k-1}; w; y)$, we infer from (6.16) that

$$|\Psi_{[k,j]}(t, w; y) - \Psi_{[k,j]}(t', w'; y')| \le A|t-t'|$$
$$+ \max\left\{\left|\Psi_{[k-1,j]}(t_{k-1}, w; y) - \Psi_{[k-1,j]}(t_{k-1}, w'; y')\right|, |y-y'|\right\} e^{1/2}.$$

Again one concludes inductively that

$$e^{1/2} \max\{|\Psi_{[k-1,j]}(t_{k-1}, w; y) - \Psi_{[k-1,j]}(t_{k-1}, w'; y')|, |y-y'|\}$$
$$\le e^{1/2} \max\left\{e^{1/2} \max\left\{|\Psi_{[k-2,j]}(t_{k-2}, w; y)\right.\right.$$
$$\left.\left. - \Psi_{[k-2,j]}(t_{k-2}, w'; y')|, |y-y'|\right\}, |y-y'|\right\}$$

$$\leq e^{(k-j-1)/2}\Big\{|\Psi_{[j+1,j]}(t_{j+1},w;y)-\Psi_{[j+1,j]}(t_{j+1},w';y')|,|y-y'|\Big\},$$

and since

$$\begin{aligned}&|\Psi_{[j+1,j]}(t_{j+1},w;y)-\Psi_{[j+1,j]}(t_{j+1},w';y')|\\&\quad=|\Phi^{\mu_{j+1}}_{w,I_{j+1}}(t_{j+1},\bar{z}_w;y)-\Phi^{\mu_{j+1}}_{w',I_{j+1}}(t_{j+1};\bar{z}_{w'};y')|\\&\quad\leq e^{1/2}\max\{|w-w'|,|y-y'|\},\end{aligned}$$

(6.17) follows.

Concerning (6.18), recall from (6.4) that $k\leq K\leq 2\|\mathbf{a}\|\widehat{T}$. Then, (6.18) follows for any $t,t'\in I_k$, $k\leq K$, from (6.17). The general case is again obtained by using the triangle inequality and inserting intermediate time-segments. This completes the proof. □

6.2.4 Pointwise Compositions

We wish to pass from compositions of global operators (integral operators) to compositions of pointwise mappings. Consider an equidistant partition of $I=[\underline{t},\bar{t}]$ with breakpoints $\tau_i=\tau_i(I,q):=\underline{t}+i|I|/q$, for some $q\in\mathbb{N}$. Let ξ_i denote the respective midpoints of the intervals $[\tau_{i-1},\tau_i]=:J_i=J_i(I,q)\subset I$, $i=1,\dots,q$, and define

$$\rho_{i,I}(t)=\rho_i(t):=\int_{\underline{t}}^{t}\chi_{J_i}(s)ds.$$

The following simple facts will be used frequently.

Lemma 6.3 *Adhering to the above notation, the following holds:*

(a) For $t\in I_k$

$$\sum_{i=1}^{q}\rho_i(t)=\sum_{i=1}^{k-1}|J_i|+t-\tau_{k-1}\leq\frac{k|I|}{q},\tag{6.19}$$

and for $t,t'\in I$

$$\sum_{i=1}^{q}|\rho_i(t)-\rho_i(t')|\leq|t-t'|.\tag{6.20}$$

(b) Assume that $v\in L_\infty(I)$ and let $v_{J_i}:=|J_i|^{-1}\int_{J_i}v(s)ds$, $i=1,\dots,q$. Then,

$$\Big|\int_{\underline{t}}^{t}v(s)ds-\sum_{i=1}^{q}\rho_i(t)v_{J_i}\Big|\leq\frac{|I|\|v\|_{L_\infty(I)}}{2q},\quad t\in I.\tag{6.21}$$

(c) Assume that $v \in \mathrm{Lip}_1(I)$ with Lipschitz constant L'. Then

$$\Big| \int_{\underline{t}}^{t} v(s)ds - \sum_{i=1}^{q} \rho_i(t) v(\xi_i) \Big| \le \frac{|I|^2 L'}{2q}. \tag{6.22}$$

The proof is elementary and given for completeness in Appendix B.

We approximate now $\Phi_{x,I}$ in a first step by the piecewise affine-in-time function

$$P_{x,I,q}(t, \bar{z}; y) := x + \sum_{i=1}^{q} \rho_i(t) \bar{\mathbf{a}}_i(\bar{z}(\xi_i); y),$$

where, depending on our hypothesis on $\mathbf{a}$, we set

$$\bar{\mathbf{a}}_i(\bar{z}; y) := \begin{cases} \mathbf{a}(\xi_i, \bar{z}; y), & \text{(A1) in case (3.3) holds,} \\ \mathbf{a}_{J_i}(\bar{z}; y) := |J_i|^{-1} \int_{J_i} \mathbf{a}(s, \bar{z}; y) ds, & \text{(A2) in case (3.3) holds.} \end{cases} \tag{6.23}$$

We record for later use that, by (6.19), the following analog to (6.7) holds

$$\begin{aligned} |P_{x,I,q}(t, w; y) - P_{x,I,q}(t, \tilde{w}; y)| &\le \sum_{i=1}^{q} \rho_i(t) |\bar{\mathbf{a}}_i(w(\xi_i); y) - \bar{\mathbf{a}}_i(\tilde{w}(\xi_i); y)| \\ &\le L \|w - \tilde{w}\|_{L_\infty(I)} \frac{i(t)|I|}{q} \le \frac{1}{2} \|w - \tilde{w}\|_{L_\infty(I)}, \end{aligned}$$

where we have used (6.4), $L \le \|\mathbf{a}\|$, and the fact that for either version of $\bar{\mathbf{a}}_i$ Lipschitz constants with respect to $\mathbb{R}^m \times \mathcal{Y}$ are preserved.

Next we estimate the deviation between $\Phi_{x,I}$ and $P_{x,I,q}$.

Lemma 6.4 *Assume that (3.2) holds and that $\bar{z} \in L_\infty(I; \mathbb{R}^m)$ satisfies*

$$\|\bar{z} - \bar{z}(\xi_i)\|_{L_\infty(J_i; \mathbb{R}^m)} \le \frac{A|I|}{2q}. \tag{6.24}$$

Then one has

$$\begin{aligned} \Big| \Phi_{x,I}(t, \bar{z}; y) - P_{x,I,q}(t, \bar{z}; y) \Big| &\le \left\{ \begin{array}{l} \frac{(1+A)L|I|^2}{2q}, \ \textit{when } \bar{\mathbf{a}}_i = \mathbf{a}(\xi_i), \\ \frac{(L|I|+1)A|I|}{2q} \ \textit{when } \bar{\mathbf{a}}_i = \mathbf{a}_{J_i}. \end{array} \right\} \\ &\le \frac{A|I|}{q} \le \frac{1}{2q}. \end{aligned} \tag{6.25}$$

The second but last inequality is relevant when $L \gg A$ so that $|I|$ is correspondingly small.

Proof : Let $\bar{z}(\underline{\xi})$ denote the piecewise constant $\bar{z}(\underline{\xi})|_{J_i} = \bar{z}(\xi_i)$ to obtain from (3.2) and (3.12)

$$\begin{aligned}
&\left|\Phi_{x,I}(t,\bar{z};y) - P_{x,I,q}(t,\bar{z};y)\right| \\
&\quad\le \left|\Phi_{x,I}(t,\bar{z};y) - \Phi_{x,I}(t,\bar{z}(\underline{\xi});y)\right| + \left|\Phi_{x,I}(t,\bar{z}(\underline{\xi});y) - P_{x,I,q}(t,\bar{z};y)\right| \\
&\quad\le |I|L\|\bar{z}(\underline{\xi}) - \bar{z}\|_{L_\infty(I;\mathbb{R}^m)} + \left|\Phi_{x,I}(t,\bar{z}(\underline{\xi});y) - P_{x,I,q}(t,\bar{z};y)\right| \\
&\quad\le \frac{LA|I|^2}{2q} + \left|\Phi_{x,I}(t,\bar{z}(\underline{\xi});y) - P_{x,I,q}(t,\bar{z};y)\right|,
\end{aligned}$$

where we have used (6.24). In case (A2), i.e., $\bar{\mathbf{a}}_i(\cdot;\cdot) = \mathbf{a}_{J_i}(\cdot;\cdot)$, (6.21) yields, in view of (3.12),

$$\left|\Phi_{x,I}(t,\bar{z}(\underline{\xi});y) - P_{x,I,q}(t,\bar{z};y)\right| \le \frac{A|I|}{2q}.$$

Thus, in this case

$$\left|\Phi_{x,I}(t,\bar{z};y) - P_{x,I,q}(t,\bar{z};y)\right| \le \frac{(L|I|+1)A|I|}{2q}.$$

Now suppose (A1), i.e., $\bar{\mathbf{a}}_i(\cdot;\cdot) = \mathbf{a}(\xi_i,\cdot;\cdot)$ under assumption (3.3). Then, we apply Lemma 6.3, (c), to $v(s) = \mathbf{a}(s,\bar{z};y)$ and, by (3.3) and (3.12), $(L' \le L)$, obtain

$$\left|\Phi_{x,I}(t,\bar{z}(\underline{\xi});y) - P_{x,I,q}(t,\bar{z};y)\right| \le \frac{L|I|^2}{2q},$$

which confirms the first inequality. On account of the assumption $1 \le A \le L$ (see (3.12), (3.13)), (6.4) ensures that the first case is bounded by $A|I|/(2q)$ while the second case is bounded by $A|I|3/(4q)$. Again (6.4) concludes the proof. □

Remark 6.3 The hypothesis (6.24) in Lemma 6.4 is valid in the following cases:

(i) $\bar{z} = \bar{z}_w$ for some $w \in \mathbb{R}^m$ on I;
(ii) $\bar{z}$ results from applying $\Phi_{x,I}$, i.e., $\bar{z}(t) = \Phi_{x,I}(t,w;y)$ for some $w \in L_\infty(I;\mathbb{R}^m)$, $y \in \mathcal{Y}$.
(iii) $\bar{z}(t) = P_{x,I,q}(t,w;y)$ results from applying $P_{x,I,q}$ to some w, y as above.

In fact, in case (i) one has $\bar{z}_w(s) - \bar{z}_w(\xi_i) = 0$. In case (ii) one has for $s \in J_i$

$$|\bar{z}(s) - \bar{z}(\xi_i)| = |\Phi_{x,I}(s,w;y) - \Phi_{x,I}(\xi_i,w;y)| \le \left|\int_{\xi_i}^{s} |\mathbf{a}(s',w(s');y)|ds'\right| \le \frac{A|I|}{2q},$$

where we have again used (3.12). Finally for (iii), we have for $s \in J_i$, by (6.20) and (3.12),

$$|\bar{z}(s) - \bar{z}(\xi_i)| = |P_{x,I,q}(s,w;y) - P_{x,I,q}(\xi_i,w;y)|$$

$$\leq \sum_{k=1}^{q} |\rho_k(s) - \rho_k(\xi_i)||\bar{\mathbf{a}}_k(w(\xi_k); y)|$$
$$\leq \frac{A|I|}{2q}.$$

which confirms the claim. □

Compositional representation of $P_{x,I,q}$: Note that $P_{\cdot,I,q}$ can be written as a composition

$$P_{x,I,q}(t, \bar{z}, y) = (g^{2,q} \circ g^{1,q})(t, x, \bar{z}; y). \tag{6.26}$$

In slight abuse of notation we identify a piecewise constant $\bar{z}$ with the vector $\bar{z}(\underline{\xi}) := (\bar{z}(\xi_1), \ldots, \bar{z}(\xi_q)) \in \mathbb{R}^{qm}$ when writing

$$g^{1,q} : \big(t, x, \bar{z}; y\big) \mapsto \big(\rho_1(t), \ldots, \rho_q(t), x, \bar{\mathbf{a}}_1(\bar{z}(\xi_1); y), \ldots, \bar{\mathbf{a}}_q(\bar{z}(\xi_q); y)\big) \in \mathbb{R}^{(q+1)m+q},$$

and the bi-linear map

$$g^{2,q} : (r_1, \ldots, r_q, x, w^1, \ldots, w^q) \mapsto x + \sum_{i=1}^{k} r_i w^i \in \mathbb{R}^m.$$

6.2.5 Dimension-Sparse Approximation

The approximation $P_{x,I,q}$ to $\Phi_{x,I}$ still involves the functions $\bar{\mathbf{a}}_i(\bar{z}; y)$ which eventually need to be approximated by finitely parametrized expressions. Here we use the structural assumptions on the convection field. In case (A1) from (6.23) $\mathbf{a} \in L_\infty([0, \widehat{T}]; \mathcal{A}^{\gamma,\mathrm{s}})$ immediately implies that $\bar{\mathbf{a}}_i(\cdot;\cdot) = \mathbf{a}(\xi_i, \cdot;\cdot)$ belong to $\mathcal{A}^{\gamma,\mathrm{s}}$, uniformly in $i = 1, \ldots, q$, $q \in \mathbb{N}$. Hence, for each $i = 1, \ldots, q$, $N \in \mathbb{N}$, there is a composition $\tilde{A}_{N,i} \in \mathfrak{C}_{N,\mathrm{s}}$ such that

$$\max_{i=1,\ldots,q} |\bar{\mathbf{a}}_i(\bar{z}; y) - \tilde{A}_{N,i}(\bar{z}; y)| \leq \gamma(N)^{-1} \|\mathbf{a}\|_{L_\infty(I;\mathcal{A}^{\gamma,\mathrm{s}})}, \quad |\!|\!|\tilde{A}_{N,i}|\!|\!|_{N,\mathrm{s}} \leq \|\mathbf{a}\|,$$

recall $\|\mathbf{a}\| := \|\mathbf{a}\|_{L_\infty(I;\mathcal{A}^{\gamma,\mathrm{s}})}$. In case (A2), the same conclusion holds, due to (3.5).

Lemma 6.5 *We adhere to the definitions* $\bar{\mathbf{a}}_i(\cdot;\cdot) = a(\xi_i, \cdot;\cdot)$ *or* $\mathbf{a}_i = \mathbf{a}_{J_i}$ *when* (3.3)*, respectively* (3.5)*, hold and let*

$$A_{x,I,q,N}(t, \bar{z}; y) := x + \sum_{i=1}^{q} \rho_i(t) \tilde{A}_{N,i}(\bar{z}; y). \tag{6.27}$$

Then, for either version of $\bar{\mathbf{a}}_i$ *one has*

$$\Big|\Phi_{x,I}(t, \bar{z}; y) - A_{x,I,q,N}(t, \bar{z}; y)\Big| \leq |I|\Big\{\frac{A}{q} + \frac{\|\mathbf{a}\|}{\gamma(N)}\Big\}.$$

In particular, choosing

$$q = q(\tau) := \left\lceil \frac{2A|I|}{\tau} \right\rceil, \quad N = N(\tau) = \lceil \gamma^{-1}(2|I|\|\mathbf{a}\|/\tau) \rceil, \tag{6.28}$$

we have

$$|\Phi_{x,I}(t, \bar{z}; y) - A_{x,I,q(\tau),N(\tau)}(t, \bar{z}; y)| \le \tau. \tag{6.29}$$

In what follows we write briefly $A_{x,I,\tau} := A_{x,I,q(\tau),N(\tau)}$ *with dimensionality vector* $\mathsf{D}_{x,I,\tau}$.

Proof : For (A1) it follows from (6.25) and Lemma 6.3, (a), (see also (3.2))

$$\begin{aligned} \left|\Phi_{x,I}(t, \bar{z}; y) - A_{x,I,q,N}(t, \bar{z}; y)\right| &\le \frac{A|I|}{q} + \sum_{i=1}^{q} \rho_i(t)|\tilde{\mathbf{a}}_i(\bar{z}; y) - \tilde{A}_{N,i}(\bar{z}; y)| \\ &\le \frac{A|I|}{q} + \frac{|I|\|\mathbf{a}\|}{\gamma(N)} = |I|\Big\{\frac{A}{q} + \frac{\|\mathbf{a}\|}{\gamma(N)}\Big\}. \end{aligned}$$

The remainder of the assertion is an obvious consequence. □

Remark 6.4

(a) Suppose that $\mathsf{D}_{i,I,q,N}$ is the dimensionality vector of $\tilde{A}_{N,i}$. Then, a corresponding realization of $A_{x,I,q,N}$ results from parallelization of the $\tilde{A}_{N,i}$. One easily concludes from Remark 2.3 that the resulting dimensionality vector $\mathsf{D}_{x,I,q,N}$ of $A_{x,I,q,N}$ is bounded by

$$\mathfrak{N}(\mathsf{D}_{x,I,q,N}) \le q \max_{i=1,\dots,q} \mathfrak{N}\big(\mathsf{D}_{i,I,q,N}\big) \le qN. \tag{6.30}$$

Thus, by (6.28), one has for an absolute constant (depending only on m)

$$\mathfrak{N}(\mathsf{D}_{x,I,\tau}) \lesssim q(\tau)\gamma^{-1}(\|\mathbf{a}\|2|I|/\tau) \le \frac{2A|I|\gamma^{-1}(2|I|\|\mathbf{a}\|/\tau)}{\tau} \le \frac{A\gamma^{-1}(1/\tau)}{\|\mathbf{a}\|\tau},$$

where we have used (6.4).

(b) $A_{\cdot,I,q,N}$ has a compositional representation analogous to (6.26), obtained by replacing $\mathbf{a}(\xi_i, \cdot, \cdot)$ by $\tilde{A}_{N,i}$. Since by assumption $\mathrm{s} \ge m$, one can see from (6.26), that s-dimension sparsity of the $\tilde{A}_{N,i}$ is inherited by the mappings $A_{\cdot,I,q,N}$ and hence by their compositions.

6.2.6 Lipschitz Continuity of Pointwise Compositions

As a final prerequisit, to eventually control the stability of compositions of $A_{x,I,q,N}$, we need bounds for the Lipschitz constants of such compositions. To that end, suppose that

$$\tilde{A}_{N,i} = (\tilde{A}_{N,i})^{n_i} \circ \cdots \circ (\tilde{A}_{N,i})^{1}, \quad i = 1, \dots, q, \tag{6.31}$$

where, by definition of $\mathfrak{C}_{N,\mathrm{S}}$, each component $(\tilde{A}_{N,i})^j_\nu$, $1 \le \nu \le d_j$ depends for $j < n_i$ only on at most S variables or is at most multi-linear.

To proceed, recall also that the Lipschitz constants of the factors in $\tilde{A}_{N,i}$ as well as the Lipschitz constants $L_{[n_i,j]}(\tilde{A}_{N,i})$ of the partial compositions $(\tilde{A}_{N,i})^{n_i} \circ \cdots \circ (\tilde{A}_{N,i})^j$ are controlled by

$$\max_{1\le i\le q} |\!|\!| \tilde{A}_{N,i} |\!|\!|_{N,\mathrm{Lip}} \le \|\mathbf{a}\| := \|\mathbf{a}\|_{L_\infty(I;\mathcal{A}^{\gamma,\mathrm{S}})}. \tag{6.32}$$

Lemma 6.6 *For any $q \in \mathbb{N}$, $k \in \mathbb{N}$, and $t \in I$, one has*

$$|A^k_{x,I,q,N}(t,\bar{z};y) - A^k_{x,I,q,N}(t,\bar{z}';y')| \le \frac{(\|\mathbf{a}\||I|)^k}{k!} \max\left\{|y-y'|, \|\bar{z}-\bar{z}'\|_{L_\infty(I;\mathbb{R}^m)}\right\},$$

for $x \in D$, $\bar{z}(\underline{\xi})$, $\bar{z}(\underline{\xi}) \in \mathbb{R}^{mq}$. Similarly, when $\bar{z} = \bar{z}_x$, $\bar{z}' = \bar{z}_{x'}$, one has for all $x, x' \in D$,

$$\begin{aligned} &|A^k_{x,I,q,N}(t,\bar{z}_x;y) - A^k_{x',I,q,N}(t',\bar{z}_{x'};y')| \\ &\quad \le \|\mathbf{a}\||t-t'| + \max\left\{|y-y'|, |x-x'|\right\} e^{\|\mathbf{a}\||I|}. \end{aligned} \tag{6.33}$$

Finally, the $A^k_{\cdot,I,q,N}$ belong to $\mathfrak{C}_{CkqN,\mathrm{S}}$, where C is a fixed constant.

The reasoning is analogous to the proof of Lemma 6.2 based on the smoothing effect of multiple integration, here in terms of multiple summation. The proof is therefore deferred to Appendinx B.

6.3 Proof of Theorem 4.1

Step 1 - construction of an ε-accurate pointwise composition: Given the ε-accurate approximation of the characteristic field by compositions of global operators Z_ε from (6.12), we construct now a *pointwise compositional* counterpart. Specifically, we define approximations $\widetilde{\Psi}_{[k,j]}$, $\widetilde{\Psi}_{\underline{\eta}^k} = \widetilde{\Psi}_{[k,0]}$ to the (global) counterparts $\Psi_{[k,j]}$, $\Psi_{\underline{\eta}^k}$. from (6.34), (6.35). We adhere to the meaning of μ_k, η_k from (6.8), and replace $\Phi_{x,I}$ by $A_{x,I,\tau} = A_{x,I,N(\tau),q(\tau)}$. Precisely, let for $j < k$, $w \in \mathbb{R}^m$, and for a new vector of tolerances

$$\underline{\tau} = (\underline{\tau}^1, \ldots, \underline{\tau}^K), \quad \text{with sections} \quad \underline{\tau}^k = (\tau_1, \ldots, \tau_k), \quad 1 \le k \le K,$$

yet to be chosen. We define for $t \in I_k$

$$\begin{aligned} &\widetilde{\Psi}_{[k,j]}(t,w;y) := A^{\mu_k}_{w_{k-1,j},I_k,\tau_k}(t,\bar{z}_{w_{k-1,j}};y), \\ &w_{k-1,j} := \widetilde{\Psi}_{[k-1,j]}(t_{k-1},w;y), \\ &\widetilde{\Psi}_{[j+1,j]}(t,w;y) := A^{\mu_{j+1}}_{w,I_k,\tau_{j+1}}(t,\bar{z}_w;y). \end{aligned} \tag{6.34}$$

We denote as before

$$\widetilde{\Psi}_{\underline{\tau}^k}(t,x;y) := \widetilde{\Psi}_{[k,0]}(t,x;y). \tag{6.35}$$

We choose $\tau_k = \tau$ all equal so that for η_k, given by (6.10),

$$|\Phi^{\mu_k}_{w,I_k}(t,\bar{z}_w;y) - A^{\mu_k}_{w,I_k,\tau_k}(t,\bar{z}_w;y)| \le \eta_k, \quad k \le K.$$

Since, by Lemma 6.2, (6.13), the Lipschitz constants $L_{[k,j]}$ of $k-j$ partial compositions of $\Phi_{w,I}$ in Remark 2.13 are bounded by $\frac{2^{-(k-j)}}{(k-j)!}$, we conclude

$$|\Phi^{\mu_k}_{w,I_k}(t,\bar{z}_w;y) - A^{\mu_k}_{w,I_k,\tau_k}(t,\bar{z}_w;y)| \le \tau + \tau\sum_{j=1}^{\mu_k-1}\frac{2^{-\mu_k-j}}{(\mu_k-j)!} \le \tau e^{1/2}.$$

On account of (6.10), choosing $\tau = \tau(\varepsilon)$ such that $\tau e^{1/2} \le \eta(\varepsilon)$ from (6.10), i.e.,

$$\tau_k(\varepsilon) = \tau(\varepsilon) = e^{-1/2}\eta(\varepsilon) = \varepsilon e^{-K/2}(1-e^{-1/2}), \quad k = 1,\dots,K, \tag{6.36}$$

yields via the same reasoning as in (6.11)

$$\|\Psi_{\underline{\eta}^k(\varepsilon)} - \widetilde{\Psi}_{\underline{\tau}^k(\varepsilon)}\|_{L_\infty(\Omega(I_k))} \le \varepsilon. \tag{6.37}$$

In summary, we obtain as before

$$\|z - \widetilde{Z}_\varepsilon\|_{L_\infty(\Omega;\mathbb{R}^m)} \le 2\varepsilon, \quad \text{where} \quad \widetilde{Z}_\varepsilon(t,x;y) := \sum_{k=1}^{K}\chi_{I_k}(t)\widetilde{\Psi}_{\underline{\tau}^k(\varepsilon)}(t,x;y). \tag{6.38}$$

Step 2 - Complexity of $\widetilde{Z}_\varepsilon$: It follows from Remark 2.3 that

$$\mathfrak{N}(\widetilde{Z}_\varepsilon) \le \sum_{k=1}^{K}\mu_k(\varepsilon)\mathfrak{N}(A_{\cdot;I_k,\tau_k(\varepsilon)}). \tag{6.39}$$

On account of (6.10) and (6.8) we have (recall $|I_k| = \widehat{T}/K$ and $K/2 = \widehat{T}\|\mathbf{a}\|$ by (6.4))

$$\mu_k(\varepsilon) = \left|\log_2\left(\frac{e^{K/2}}{2(e^{1/2}-1)\varepsilon}\right)\right| \eqsim \left|\log_2\left(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\right)\right|. \tag{6.40}$$

Furthermore, (6.36) in conjunction with Remark 6.4 yields (for the range of γ under consideration, see (2.25))

$$\mathfrak{N}(A_{\cdot;I_k,\tau_k(\varepsilon)}) \eqsim \frac{Ae^{K/2}}{\|\mathbf{a}\|\varepsilon}\gamma^{-1}\Big(\frac{e^{K/2}}{(1-e^{-1/2})\varepsilon}\Big) \eqsim \frac{Ae^{\|\mathbf{a}\|\widehat{T}}}{\|\mathbf{a}\|\varepsilon}\gamma^{-1}\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\Big).$$

Substituting this into (6.39), yields

$$\mathfrak{N}(\widetilde{Z}_\varepsilon) \eqsim K\left|\log_2\left(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\right)\right|\frac{Ae^{\|\mathbf{a}\|\widehat{T}}}{\|\mathbf{a}\|\varepsilon}\gamma^{-1}\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\Big)$$

$$\eqsim A\widehat{T}\Big|\log_2\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\Big)\Big|\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\gamma^{-1}\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\Big).$$

By (2.25), we obtain

$$\gamma^{-1}\big(e^{\|\mathbf{a}\|\widehat{T}}\varepsilon^{-1}\big)\eqsim\begin{cases}C_a^{-1/\alpha}e^{\|\mathbf{a}\|\widehat{T}/\alpha}\varepsilon^{-\frac{1}{\alpha}}, & \text{for } \gamma\sim\ (\text{alg}),\\ \frac{1}{\alpha}\ln\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{C_e\varepsilon}\Big), & \text{for } \gamma\sim\ (\text{exp}).\end{cases}$$

We conclude that for $\gamma\sim(\text{alg})$, (see (2.24))

$$\mathfrak{N}(\widetilde{Z}_\varepsilon)\eqsim AC_a^{-1/\alpha}\widehat{T}\log_2\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\Big)\Big(\frac{e^{\widehat{T}\|\mathbf{a}\|}}{\varepsilon}\Big)^{\frac{\alpha+1}{\alpha}}:=\phi_{\text{alg}}\big(e^{\|\mathbf{a}\|\widehat{T}}/\varepsilon\big).\tag{6.41}$$

For exponential growth $\gamma\sim(\text{exp})$ (since $\|\mathbf{a}\|\geq L\geq 1$, (3.12)) we obtain

$$\mathfrak{N}(\widetilde{Z}_\varepsilon)\eqsim\frac{1}{\alpha}A\widehat{T}\Big(\log_2\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\Big)\Big)^2\Big(\frac{e^{\widehat{T}\|\mathbf{a}\|}}{\varepsilon}\Big):=\phi_{\text{exp}}\big(e^{\|\mathbf{a}\|\widehat{T}}/\varepsilon\big),\tag{6.42}$$

provided that $C_e\gtrsim 1$.

Remark 6.5 On account of Remark 6.4, (b) and Lemma 6.6, we conclude that $\widetilde{Z}_\varepsilon\in\mathfrak{C}_{CN_\varepsilon,\mathrm{S}}$ for some uniform constant C and $\|\widetilde{Z}_\varepsilon\|_{N_\varepsilon,\mathrm{S}}\leq e^{\|\mathbf{a}\|\widehat{T}}$ for N_ε, defined by the respective right hand sides in (6.41), (6.42).

Step 3 - Convergence rates: To determine the convergence rates, corresponding to (6.12), we apply Remark 6.1 and Lemma 6.1 to $N_\varepsilon=\phi_{\text{alg/exp}}\Big(e^{\|\mathbf{a}\|\widehat{T}}/\varepsilon\Big)$ from (6.41) and (6.42). (4.4) follows then by straightforward calculations.

The following statement follows now from Remark 6.1 and the above observations.

Remark 6.6 In summary we have shown that for each $N\in\mathbb{N}$ there exists an S-dimension-sparse compositional representation $\widetilde{Z}_N$ satisfying

$$\|z-\widetilde{Z}_N\|_{L_\infty(\Omega;\mathbb{R}^m)}\lesssim e^{\|\mathbf{a}\|\widehat{T}}\widetilde{\gamma}(N)^{-1},\quad N\in\mathbb{N},$$

with $\widetilde{\gamma}$ from (4.1). This confirms (4.2).

Step 4 - Stability of $\widetilde{Z}_\varepsilon$: It follows from Lemma 6.6, (6.33), that the Lipschitz constants of partial compositions of the approximations $\widetilde{\Psi}_{\underline{\tau}^k}$ from (6.35) remain uniformly bounded by $\|\mathbf{a}\|+e^{\|\mathbf{a}\||I_k|}$. Specifically,

$$|\widetilde{Z}_\varepsilon(t,x,y)-\widetilde{Z}_\varepsilon(t,x'y')|\leq\max\{|x-x'|,|y-y'|\}e^{\|a\|\widehat{T}},\quad t\in[0,\widehat{T}],$$

i.e.,

$$\|\widetilde{Z}_\varepsilon(t,\cdot,\cdot)\|_{N_\varepsilon,\mathrm{S}}\leq e^{\|\mathbf{a}\|\widehat{T}},\quad t\in[0,\widehat{T}].$$

Moreover, the growth functions $\hat{\gamma}$ satisfy $\hat{\gamma}(N_\varepsilon)\eqsim\varepsilon$. By Remark 6.1, this finishes the proof of Theorem 4.1. □

6.4 Proof of Theorem 4.2

The main step is to invoke Lemma 2.1 and Proposition 2.14 to approximate each $\tilde{A}_{N,i}(\bar{z}, y) \in \mathfrak{C}_{N,\mathrm{S}}$ in (6.27) by a DNN. In essence we follow the same steps as in the preceding section.

Approximation of data—accuracy: Recall from (6.31) that $\tilde{A}_{N,i}$ has a compositional representation $\tilde{A}_{N,i} = (\tilde{A}_{N,i})^{n_i} \circ \cdots \circ (\tilde{A}_{N,i})^1$ of some depth $n_i \le \mathfrak{N}(\tilde{A}_{N,i})$.

For a given tolerance $\delta > 0$, we construct next a network $\mathcal{N}_{i,\delta,N,q}(\bar{z}, y)$ approximating $\tilde{A}_{N,i}$ with accuracy $\|\mathbf{a}\|\delta$. Specifically, we approximate each component $(\tilde{A}_{N,i})_\nu^j$, $\nu = 1, \ldots, d_j$, in the jth composition factor by a δ/n_i-accurate neural network $(\mathcal{N})_\nu^j$. One infers from Lemma 2.1, (2.26), (6.32), that (since $\|\mathbf{a}\| \ge 1$)

$$\begin{aligned}\|\tilde{A}_{N,i} - \mathcal{N}_{i,\delta,N,q}\|_{L_\infty(\mathbb{R}^m \times \mathcal{Y};\mathbb{R}^m)} &\le n_i^{-1}\Big\{\delta + \sum_{r=1}^{n_i-1} \delta L_{[n_i,r+1]}(\tilde{A}_{N,i})\Big\} \\ &\le n_i^{-1}\Big\{\delta + \|\mathbf{a}\| \sum_{r=1}^{n_i-1} \delta\Big\} \\ &\le \|\mathbf{a}\|\delta, \quad i = 1, \ldots, q. \end{aligned} \tag{6.43}$$

On the other hand, we invoke (2.27) in Lemma 2.1 to conclude

$$\#\mathcal{N}_{i,\delta,N,q} \le C_{\mathrm{S}} \|\mathbf{a}\|^{\mathrm{S}} N n_i^{\mathrm{S}} \delta^{-\mathrm{S}} \big|\, |\log_2 \delta| + \log_2 n_i \,\big|, \quad i = 1, \ldots, q, \tag{6.44}$$

where the constant C_{S} depends only on S. Now we define in analogy to (6.27)

$$\mathcal{N}_{x,I,q,N,\delta}(t, z; y) := x + \sum_{i=1}^{q} \rho_i(t) \mathcal{N}_{i,\delta,N,q}(z, y), \quad t \in I, \tag{6.45}$$

where we recall $I = [\underline{t}, \bar{t}]$, $J_i = J_i(I, q) := \underline{t} + \big[\frac{(i-1)|I|}{q}, \frac{i|I|}{q}\big]$ and $\rho_i(t) := \int_{\underline{t}}^{t} \chi_{J_i}(s) ds$. Taking

$$\mathcal{N}_{x,I,\tau}(t; z; y) := \mathcal{N}_{x,I,q(\tau/2),N(\tau/2),\delta(\tau)}(t; z; y),$$

with $q(\tau/2)$, $N(\tau/2)$, defined according to (6.28), and

$$\delta = \delta(\tau) := \frac{\tau}{2\|\mathbf{a}\|},$$

we conclude that $\|A_{x,I,\tau} - \mathcal{N}_{x,I,\tau}\|_{L_\infty(\Omega(I);R^m)} \le \tau/2$. Hence, by (6.29),

$$\Big|\Phi_{x,I}(t, z; y) - \mathcal{N}_{x,I,\tau}(t, z; y)\Big| \le \tau.$$

To estimate the complexity of $\mathcal{N}_{x,I,\tau}$ we use (6.44) and bound the depths n_i by $N(\tau/2) \sim \gamma^{-1}\big(4|I|\|\mathbf{a}\|/\tau\big) = \gamma^{-1}\big(2/\tau\big)$. Arguing as in (6.30) and recalling (6.28), we then have (since $q(\tau) = 2A|I|/\tau = \frac{A}{\|\mathbf{a}\|\tau}$)

$$\begin{aligned}
\#\mathcal{N}_{x,I,\tau} &\eqsim q(\tau/2) \max_{i=1,\dots,q(\tau/2)} \#\mathcal{N}_{i,\delta(\tau/2),N(\tau/2)} \\
&\eqsim \frac{2A}{\|\mathbf{a}\|\tau}\|\mathbf{a}\|^{\mathrm{s}}\gamma^{-1}(2/\tau)^{1+\mathrm{s}}\delta(\tau)^{-\mathrm{s}}\big|\log_2\delta(\tau)| + \log_2 N(\tau/2)\big| \\
&\eqsim A2^{\mathrm{s}}\|\mathbf{a}\|^{2\mathrm{s}-1}\tau^{-(1+\mathrm{s})}\gamma^{-1}(2/\tau)^{1+\mathrm{s}}\Big|\log_2\frac{2\|\mathbf{a}\|\gamma^{-1}(2/\tau)}{\tau}\Big|.
\end{aligned}$$

For algebraic growth $\gamma(r) = C_a r^\alpha$ we obtain $\gamma^{-1}\big(2/\tau\big) = (2/C_a)^{1/\alpha}\tau^{-1/\alpha}$, and $\log_2\gamma^{-1}(2/\tau) \sim \alpha^{-1}\log_2(2/C_a\tau)$, so that $\Big|\log_2\frac{2\|\mathbf{a}\|\gamma^{-1}(2/\tau)}{\tau}\Big| = \frac{\alpha+1}{\alpha}\Big|\log_2\frac{2}{\tau}\Big(\frac{\|\mathbf{a}\|}{C_a}\Big)^{\frac{\alpha}{\alpha+1}}\Big|$. For $\gamma\sim$(exp), we have $\Big|\log_2\frac{2\|\mathbf{a}\|\gamma^{-1}(2/\tau)}{\tau}\Big| = \log_2\frac{2\|\mathbf{a}\|\ln(2/C_e\tau)}{\alpha\tau}$. Thus,

$$\#\mathcal{N}_{x,I,\tau} \eqsim A2^{\mathrm{s}}\|\mathbf{a}\|^{2\mathrm{s}-1}\begin{cases} C_a^{-1/\alpha}\tau^{-\frac{(1+\alpha)(1+\mathrm{s})}{\alpha}}|\log_2\tau|, & \text{when } \gamma\sim\ (\text{alg}), \\ \alpha^{-(\mathrm{s}+1)}\tau^{-(\mathrm{s}+1)}|\log_2\tau|^{2+\mathrm{s}}, & \text{when } \gamma\sim\ (\text{exp}), \end{cases} \tag{6.46}$$

where we recall that $C_e \gtrsim 1$, accepting a logarithmic dependence of the proportionality constant on $\|\mathbf{a}\|$ (or assume that $\tau \le \tau_0(\|\mathbf{a}\|)$).

We can now define $\widehat{\Psi}_{[k,j]}(t,w;y)$ in analogy to (6.34) with $A_{w,I,\tau_k}(t,\bar{z}_w;y)$ replaced by $\mathcal{N}_{w,I,\tau_k}(t,\bar{z}_w;y)$ and likewise $\widehat{\Psi}_{\underline{\tau}^k} = \widehat{\Psi}_{[k,0]}$ in analogy to (6.35), with the same tolerances $\tau_k(\varepsilon) = \tau(\varepsilon)$, given by (6.36). The same reasoning as in Sect. 6.3 yields (see (6.37)

$$\|\Psi_{\underline{\eta}^k(\varepsilon)} - \widehat{\Psi}_{\underline{\tau}^k(\varepsilon)}\|_{L_\infty(\Omega(I_k))} \le \varepsilon, \quad k = 1,\dots,K,$$

and hence

$$\|z - \mathcal{N}_\varepsilon\|_{L_\infty(\Omega;\mathbb{R}^m)} \le 2\varepsilon, \quad \mathcal{N}_\varepsilon(t,x;y) := \sum_{k=1}^K \chi_{I_k}(t)\widehat{\Psi}_{\underline{\tau}^k(\varepsilon)}(t,x;y).$$

Complexity: It remains to bound $\#\mathcal{N}_\varepsilon$. Invoking Remark 2.3 as before, one obtains from (6.46) with $\mu_k(\varepsilon)$ from (6.40) (see also (6.8), (6.10))

$$\begin{aligned}
\#\mathcal{N}_\varepsilon &\eqsim \sum_{k=1}^K \mu_k(\varepsilon)\mathcal{N}_{\cdot,I_k,\tau_k} \eqsim K\log_2\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{2\|a\|\varepsilon}\Big)\#\mathcal{N}_{\cdot,I,\tau(\varepsilon)} \\
&\eqsim A2^{\mathrm{s}}\|\mathbf{a}\|^{2\mathrm{s}}\widehat{T}\log_2\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{2\|\mathbf{a}\|\varepsilon}\Big) \\
&\quad\times\begin{cases} C_a^{-1/\alpha}\tau(\varepsilon)^{-\frac{(1+\alpha)(1+\mathrm{s})}{\alpha}}|\log_2\tau(\varepsilon)|, & \gamma\sim\ (\text{alg}), \\ \alpha^{-(\mathrm{s}+1)}\tau(\varepsilon)^{-(\mathrm{s}+1)}|\log_2\tau(\varepsilon)|^{2+\mathrm{s}}, & \gamma\sim\ (\text{exp}). \end{cases}
\end{aligned} \tag{6.47}$$

By (6.36),

$$C_a^{-1/\alpha}\tau(\varepsilon)^{-\frac{(1+\alpha)(1+s)}{\alpha}}|\log_2\tau(\varepsilon)| \eqsim C_a^{-1/\alpha}\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\Big)^{\frac{(1+\alpha)(1+s)}{\alpha}}\log_2\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\Big),$$

while

$$\alpha^{-(s+1)}\tau(\varepsilon)^{-(s+1)}|\log_2\tau(\varepsilon)|^{2+s} \eqsim \alpha^{-(s+1)}\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\Big)^{-(1+s)}\Big|\log_2\Big(\frac{e^{\|\mathbf{a}\|\widehat{T}}}{\varepsilon}\Big)\Big|^{2+s}.$$

Inserting these estimates into (6.47), confirms (4.3). □

The convergence rates stated in Corollary 4.3 follow now in the same way as before by applying Remark 6.1 and Lemma 6.1 to the bounds on $\mathfrak{N}(\mathcal{N}_\varepsilon)$, given in (4.3).

Regarding the stability of the networks $\mathcal{N}_\varepsilon$, there is a principal obstacle related to the fact that the precise dimension-vectors of the compositions $\tilde{A}_{N,i} = \tilde{A}^{n_i}_{N,i} \circ \cdots \circ \tilde{A}^1_{N,i}$, especially their depths n_i, are not known. Although, the Lipschitz constants of the implanted networks in each factor $\tilde{A}^v_{N,i}$ are controlled by $\|\mathbf{a}\|$ it is not clear whether the Lipschitz constants of their compositions also remain controlled by $\|\mathbf{a}\|$. Such network approximations exist by Proposition 2.14 but need no longer be s-dimension sparse. So, the only guaranteed general bound for the Lipschitz constants of the partial compositions is, in view of Proposition 2.14, $L_{[n_i,j+1]}(\mathcal{N}_{i,\delta,N,q}) \leq (c_3(1+A)\|\mathbf{a}\|)^{n_i-j}$. Hence,

$$L_{N,\delta} := \max_{j<n_i} L_{[n_i,j+1]}(\mathcal{N}_{i,\delta,N,q}) \leq (c_3(1+A)\|\mathbf{a}\|)^N, \tag{6.48}$$

where we have applied the (perhaps too pessimistic) bound $n_i \leq N$. If on the other hand, the depths n_i remain uniformly bounded by $\bar{n}$, say, one obtains a uniform Lipschitz-bound $L_{N,\delta} \leq (c_3(1+A)\|\mathbf{a}\|)^{\bar{n}}$.

Lemma 6.7 *The network approximations $\mathcal{N}_{\cdot,I,q,N,\delta}$ from (6.45) have the following Lipschitz continuity properties: for (t,x,y), $(t',x',y') \in \Omega(I)$, $\bar{z},\bar{z} \in L_\infty(I;\mathbb{R}^m)$:*

$$\begin{aligned}&|\mathcal{N}^\ell_{x,I,q,N,\delta}(t,\bar{z};y) - \mathcal{N}^\ell_{x,I,q,N,\delta}(t,\bar{z}';y')|\\ &\quad\leq \frac{(L_{N,\delta}|I|)^\ell}{\ell!}\big\{|y-y'|, \|\bar{z}-\bar{z}'\|_{L_\infty(I;\mathbb{R}^m)}\big\},\end{aligned} \tag{6.49}$$

where $L_{N,\delta}$ is given by (6.48). Similarly, when $\bar{z} = \bar{z}_x$, $\bar{z}' = \bar{z}_{x'}$, one has

$$\begin{aligned}&|\mathcal{N}^\ell_{x,I,q,N,\delta}(t,\bar{z}_x;y) - \mathcal{N}^\ell_{x',I,q,N,\delta}(t',\bar{z}_{x'},;y')|\\ &\quad\leq (1+\delta)\|\mathbf{a}\||t-t'| + \max\big\{|y-y'|,|x-x'|\big\}e^{L_{N,\delta}|I|}.\end{aligned} \tag{6.50}$$

Proof : Recall from (6.45) that $\mathcal{N}_{x,I,q,N,\delta}(t,\bar{z};y) := x + \sum_{i=1}^q \rho_i(t)\mathcal{N}_{i,\delta,N,q}(\bar{z},y)$. Then, we have for $i = 1,\ldots,q$,

$$|\mathcal{N}_{i,\delta,N,q}(\bar{z},y) - \mathcal{N}_{i,\delta,N,q}(\bar{z}',y')|$$

$$\leq |\mathcal{N}_{i,\delta,N,q}|_{\mathrm{Lip}_1(\mathbb{R}^m\times\mathcal{Y})}\max\{\|\bar z-\bar z'\|_{L_\infty(I;\mathbb{R}^m)},|y-y'|\}.$$

By the comments preceding the lemma, we obtain

$$\begin{aligned}&\left|\mathcal{N}_{x,I,q,N,\delta}(t,\bar z;y)-\mathcal{N}_{x,I,q,N,\delta}(t,\bar z';y')\right|\\&\quad\leq\sum_{i=1}^q\rho_i(t)L_{N,\delta}\max\{\|\bar z-\bar z'\|_{L_\infty(I;\mathbb{R}^m)},|y-y'|\}.\end{aligned}$$

Hence, we are in the same situation as in (6.77). Finally,

$$\begin{aligned}&\left|\mathcal{N}_{x,I,q,N,\delta}(t,\bar z;y)-\mathcal{N}_{x,I,q,N,\delta}(t',\bar z;y)\right|\\&\quad\leq\|\mathbf{a}\||t-t'|+\sum_{i=1}^q|\rho_i(t)-\rho_i(t')||\bar{\mathbf{a}}_i(\xi_i;y)-\mathcal{N}_{i,\delta,N,q}(\bar z(\xi_i);y)|\\&\quad\leq(1+\delta)\|\mathbf{a}\||t-t'|,\end{aligned}$$

where we have used (6.43). Therefore, the claim follows by the same arguments as used in the proof of Lemma 6.6. □

Regarding Remark 4.5, recall that $N(\varepsilon)\sim\gamma^{-1}(2/\tau(\varepsilon))$, where $\tau(\varepsilon)\asymp\varepsilon e^{-\|\mathbf{a}\|\widehat{T}}$, $\delta(\varepsilon)=\tau(\varepsilon)/2\|\mathbf{a}\|$ so that by (6.48),

$$L_{N(\varepsilon),\delta(e)}\asymp\begin{cases}(c_3(1+A)\|\mathbf{a}\|)^{(2/C_a)^{\frac{1}{\alpha}}\varepsilon^{-1/\alpha}e^{\|\mathbf{a}\|\widehat{T}/\alpha}}, & \text{in case } \gamma\sim\ (\mathrm{alg}),\\ (c_3(1+A)\|\mathbf{a}\|)^{\frac{1}{\alpha}\left(|\ln\varepsilon|+\|\mathbf{a}\|\widehat{T}\right)}, & \text{in case } \gamma\sim\ (\exp).\end{cases}$$

This confirms Remark 4.5. □

6.5 Proof of Corollary 4.6 and Theorem 4.7

We first prove Corollary 4.6. To apply Theorem 4.2 and Corollary 4.3 note first that condition (3.5) is applicable, i.e., version (A2) can be used. In fact, $\mathbf{a}_j(t;\cdot)=\omega_j\mathbf{a}_j^\circ(t,\cdot)\in\mathrm{Lip}_1(\mathbb{R}^m)$, uniformly in $t\in[0,\widehat{T}]$, immediately implies that $\bar{\mathbf{a}}_{j,i}=\mathbf{a}_{j,J_i}=\omega_j\mathbf{a}^\circ_{j,J_i}$ belongs to $\mathrm{Lip}_1(\mathbb{R}^m)$, for $i=1,\dots,q$, with the same Lipschitz constants $\Lambda\omega_j$ from (3.13).

Next we recall from Remark 3.2 that $\mathbf{a}\in\mathfrak{C}_{N_\mathbf{a},m}$ with $N_\mathbf{a}=1+d_y(1+m^2)$. Hence, the simplest compositional approximations $A_N(t,\cdot)$ to $\mathbf{a}$ is

$$A_N(t,x;y)=\begin{cases}0, & N<N_\mathbf{a},\\ \mathbf{a}(t,x;y), & N\geq N_\mathbf{a}.\end{cases}$$

In view of Remark 3.2, (3.11), one obtains for $\gamma(r)=C_e e^{\alpha r}$ and all $t\in[0,\widehat{T}]$,

$$\|\mathbf{a}\|_{L_\infty(\widehat{I};\mathcal{A}^{\gamma,m})}\leq\max_{N\in\mathbb{N}}\gamma(N)\Big\{\|\mathbf{a}(t)-A_N(t)\|_{L_\infty(\mathbb{R}^m\times\mathcal{Y};\mathbb{R}^m)}+\gamma(N)^{-1}|\!|\!|A_N(t)|\!|\!|_{N,m}\Big\}$$

$$\leq \begin{cases} AC_e e^{\alpha N} + (A + \Lambda|\underline{\omega}|_1), & N < N_{\mathbf{a}}, \\ A + \Lambda|\underline{\omega}|_1, & N \geq N_{\mathbf{a}}. \end{cases}$$

Taking $\alpha := N_{\mathbf{a}}^{-1}$, $C_e := 1$, yields $\|\mathbf{a}\| \leq \widehat{L} := 2A + \Lambda|\underline{\omega}|_1$. By (3.10), one has $\alpha \eqsim d_y^{-1}$ (with m-dependent proportionality). Theorem 4.1 yields then $z \in L_\infty(\widehat{I}; \mathcal{A}^{\widetilde{\gamma},m})$ with

$$\widetilde{\gamma}(r) \eqsim \frac{r}{d_y A \widehat{T}} \left| \log_2 \frac{r}{d_y A \widehat{T}} \right|^{-2}, \quad \|z\|_{L_\infty([0,\widehat{T}];\mathcal{A}^{\widetilde{\gamma},m})} \leq e^{(2A+\Lambda|\underline{\omega}|_1)\widehat{T}},$$

and hence (4.5). Now (4.6) and the expression for F follow from Corollary 4.3. □

We now turn to the proof of Theorem 4.7 approximating $\Phi_{x,I}$ in a first step by

$$P_{x,I,q}(t, \bar{z}; y) := x + \sum_{j=1}^{d_y} y_j \sum_{i=1}^{q} \rho_i(t) \mathbf{a}_{j,J_i}(\bar{z}(\xi_i)), \quad \mathbf{a}_{j,J_i}(\cdot) := \frac{\omega_j}{|J_i|} \int_{J_i} \mathbf{a}_j^\circ(s, \cdot) ds,$$

recalling that $J_i = J_i(I, q) = \underline{t} + \left[\frac{(i-1)|I|}{q}, \frac{i|I|}{q}\right]$ and $\rho_i(t) = \int_{J_i} \chi_{s \leq t}(s) ds$. For L from (3.9) and $\bar{z}$ as in Lemma 6.4, we infer from (6.25) that $|\Phi_{x,I}(t, \bar{z}; y) - P_{x,I,q}(t, \bar{z}; y)| \leq \frac{A|I|}{q}$. Invoking Proposition 2.14, we approximate the low-dimensional functions $\mathbf{a}_{j,J_i}^\circ(\cdot)$ by finitely parametrized functions such as neural networks. Specifically, there exist networks (suppressing the reference to I) $\mathcal{N}_{j,i,\delta}$ of depth $\lesssim \log_2 \delta^{-1}$ such that for $j = 1, \ldots, d_y$, $i = 1, \ldots, q$,

$$\|\mathbf{a}_{j,J_i}^\circ - \mathcal{N}_{j,i,\delta}\|_{L_\infty(\mathbb{R}^m;\mathbb{R}^m)} \leq \delta, \quad \#\mathcal{N}_{j,i,\delta} \lesssim \Lambda^m \delta^{-m} |\log_2 \delta|. \tag{6.51}$$

Then

$$\mathcal{N}_{x,I,q,\delta}(t, \bar{z}; y) := x + \sum_{j=1}^{d_y} y_j \sum_{i=1}^{q} \rho_i(t) \omega_j \mathcal{N}_{j,i,\delta}(\bar{z}(\xi_i)), \tag{6.52}$$

is indeed an m-dimension-sparse neural network. To that end, we keep viewing t as a parameter and the input-variables x, y are passed across layers, formally in a "skip-connection" format. Thus, formally we have

$$\mathcal{N}_{x,I,q,\delta}(t, \bar{z}, y) = (G_2 \circ G_1)(t, x, \bar{z}, y).$$

For better readability the following representation groups variables in a formally incorrect way and should be viewed as a t-dependent mapping into $\mathbb{R}^{1+m+d_y+mqd_y}$

$$G_1 : (t, x, \bar{z}, y) \mapsto \begin{pmatrix} x, y, \rho_1(t), \ldots, \rho_q(t) \\ \omega_1 \mathcal{N}_{1,1,\delta}(\bar{z}), \ldots \omega_{d_y} \mathcal{N}_{d_y,1,\delta}(\bar{z}), \\ \vdots \\ \omega_1 \mathcal{N}_{1,q,\delta}(\bar{z}), \ldots, \omega_{d_y} \mathcal{N}_{d_y,q,\delta}(\bar{z}) \end{pmatrix} \in \mathbb{R}^{q+m+d_y+mqd_y}, \tag{6.53}$$

which is obviously m-dimension sparse. Hence G_1 itself is a neural network whose depth is bounded by $\log_2 \delta^{-1}$. The tri-linear factor G_2 reads then

$$G_2 : (x, y, r_1, \dots, r_q, \zeta^{1,1}, \dots, \zeta^{d_y,q}) \mapsto \Big(x + \sum_{i=1}^{q} r_i \sum_{j=1}^{d_y} y_j \zeta^{j,i}\Big) \in \mathbb{R}^m.$$

Assessing the accuracy of $\mathcal{N}_{x,I,q,\delta}$ follows in essence the same lines as before. In view of (6.25) and (6.51),

$$\begin{aligned}|\Phi_{x,I}(t, \bar{z}; y) - \mathcal{N}_{x,I,q,\delta}(t, \bar{z}; y)| &\le |\Phi_{x,I}(t, \bar{z}; y) - P_{x,I,q}(t, \bar{z}; y)| \\ &+ \sum_{j=1}^{d_y} |y_j| \sum_{i=1}^{q} \rho_i(t)\omega_j |\mathcal{N}_{j,i,\delta}(\bar{z}(\xi_i)) - \mathbf{a}^{\circ}_{j,J_i}(\bar{z}(\xi_i))| \le \frac{A|I|}{q} + |\underline{\omega}|_1 |I| \delta.\end{aligned}$$

Thus, given any target tolerance $\tau > 0$, choosing

$$q(\tau) = \left\lceil \frac{2A|I|}{\tau} \right\rceil, \quad \delta(\tau) = \frac{\tau}{2|I||\underline{\omega}|_1}, \tag{6.54}$$

and abbreviating $\mathcal{N}_{x,I,\tau} := \mathcal{N}_{x,I,q(\tau),\delta(\tau)}$, we obtain

$$|\Phi_{x,I}(t, \bar{z}, y) - \mathcal{N}_{x,I,\tau}(t, \bar{z}; y)| \le \tau, \quad x \in \mathbb{R}^m, \ \bar{z} \in L_\infty(I; \mathbb{R}^m), \ t \in I.$$

Regarding the complexity of $\mathcal{N}_{x,I,\tau}$ we see from (6.53) that $\mathfrak{N}(G_1) = m + d_y + q + q d_y m^2$, $\mathfrak{N}(G_2) = 1$ because of bilinearity. Thus, $\mathfrak{N}(G_2 \circ G_1) \eqsim m^2 q d_y$ so that by (6.51) and (6.54),

$$\begin{aligned}\#\mathcal{N}_{x,I,\tau}(t, \cdot, \cdot) &\lesssim m^2 d_y q(\tau) \Lambda^m \delta(\tau)^{-m} |\log_2 \delta(\tau)| \\ &\lesssim m^2 d_y \frac{2A|I|}{\tau} \frac{(2|I|\Lambda|\underline{\omega}|_1)^m}{\tau^m} \Big| \log_2 \frac{2|I||\underline{\omega}|_1}{\tau} \Big|, \quad t \in I. \end{aligned} \tag{6.55}$$

In view of (3.11), the earlier role of $\|\mathbf{a}\|$ is now played by (see (3.9)) $\|\mathbf{a}\|_{N_\mathbf{a},m} = A + \Lambda|\underline{\omega}|_1 =: L$, and since by (6.4), $|I|L = \frac{1}{2}$, we obtain from (6.55)

$$\#\mathcal{N}_{x,I,\tau}(t, \cdot, \cdot) \lesssim m^2 d_y A |I| \tau^{-(m+1)} \Big| \log_2 \frac{2|I||\underline{\omega}|_1}{\tau} \Big|. \tag{6.56}$$

Thus, with the same number $\mu = \mu(\eta) \ge |\log_2(2\eta)^{-1}|$ from (6.8) we get $|z(t, x; y) - \Phi^{\mu}_{x,I}(t, \bar{z}_x; y)| \le \eta$. Hence, the same Z_ε, defined by (6.12), based on tolerances $\eta(\varepsilon)$ from (6.10), provide ε-accuracy of time-catenated iterates of Φ_{x,I_k} where the number K of macro-time-steps still equals $2\widehat{T}L$. We can therefore choose the same vectors of tolerances $\underline{\tau}^k(\varepsilon)$ from (6.36) (i.e., $\tau(\varepsilon) \eqsim \eta(\varepsilon) \eqsim \varepsilon e^{-K/2}$) as well as tolerances $\delta(\eta(\varepsilon)) \eqsim \varepsilon e^{-K/2}/(2|I||\underline{\omega}|_1)$. We then define $\mathcal{N}_\varepsilon$ in complete analogy to $\widetilde{Z}_\varepsilon$ from (6.38), with $A_{w,I,\tau(\varepsilon)}$ replaced by $\mathcal{N}_{w,I,\tau(\varepsilon)}$ to obtain $\|z - \mathcal{N}_\varepsilon\|_{L_\infty(\Omega;\mathbb{R}^m)} \le 2\varepsilon$.

Hence, on account of (6.56), since $|I|K = \widehat{T}$, $K/2 = L\widehat{T}$

$$\begin{aligned}
\#\mathcal{N}_\varepsilon &\lesssim m^2 d_y K \mu(\eta(\varepsilon)) A|I|\tau(\varepsilon)^{-(m+1)} \Big|\log_2 \frac{2|I||\underline{\omega}|_1}{\tau(\varepsilon)}\Big| \\
&\le A\widehat{T} m^2 d_y \tau^{-(m+1)} \Big|\log_2 \frac{2|I||\underline{\omega}|_1}{\tau(\varepsilon)}\Big| \\
&\lesssim A\widehat{T} m^2 d_y \Big(\frac{e^{L\widehat{T}}}{\varepsilon}\Big)^{m+1} \Big|\log_2 \frac{2|I|e^{L\widehat{T}}|\underline{\omega}|_1}{\varepsilon}\Big| \\
&\lesssim A\widehat{T} m^2 d_y \Big(\frac{e^{L\widehat{T}}}{\varepsilon}\Big)^{m+1} \Big|\log_2 \frac{e^{L\widehat{T}}}{\varepsilon}\Big| \\
&=: \phi\Big(\frac{e^{L\widehat{T}}}{\varepsilon}\Big),
\end{aligned}$$

where we have used that, by (6.4), $1 \lesssim |I|\Lambda|\underline{\omega}|_1 \le 1/2$. This confirms (4.8).

We apply Remark 6.1 and Lemma 6.1 to conclude that (surpressing a logarithmic dependence on $A\widehat{T}m^2$)

$$\widetilde{\gamma}(r) := (A\widehat{T}m^2)^{-\frac{1}{m+1}} \Big(\frac{r}{d_y}\Big)^{\frac{1}{m+1}} \Big|\log_2 \frac{r}{d_y}\Big|^{-\frac{2}{m+1}}$$

satisfies $\widetilde{\gamma}(\phi(s)) \eqsim s$. Hence, there exixts a network $\mathcal{N}_N$ with $\#\mathcal{N}_N \le N$ such that

$$\|z - \mathcal{N}_N\|_{L_\infty(\Omega;\mathbb{R}^m)} \lesssim e^{L\widehat{T}} \widetilde{\gamma}(N)^{-1}, \quad N \in \mathbb{N},$$

where the constant depends only on m. This proves (4.9).

It remains to estimate $\|\mathcal{N}_\varepsilon\|_{N_\varepsilon,m}$ where $N_\varepsilon := \#\mathcal{N}_\varepsilon$.

Lemma 6.8 *For $\mathcal{N}_{\cdot,I,q,\delta}(t,\cdot;\cdot)$ from (6.52) the following statements hold: Let*

$$L_\delta := \big(A + (\bar{\Lambda} + \delta)|\underline{\omega}|_1\big), \quad \bar{\Lambda} := c_3(1 + A^\circ)\Lambda, \tag{6.57}$$

with c_3 from (2.19) (see also (3.8)). Then, for (t,x,y), $(t',x',y') \in \Omega(I)$,

$$\begin{aligned}
&|\mathcal{N}^\ell_{x,I,q,\delta}(t,\bar{z};y) - \mathcal{N}^\ell_{x,I,q,\delta}(t,\bar{z}';y')| \\
&\quad\le \frac{(L_\delta|I|)^\ell}{\ell!}\big\{|y-y'|, \|\bar{z}-\bar{z}'\|_{L_\infty(I;\mathbb{R}^m)}\big\}.
\end{aligned} \tag{6.58}$$

Similarly, when $\bar{z} = \bar{z}_x$, $\bar{z}' = \bar{z}_{x'}$, one has for all $x, x' \in \mathbb{R}^m$, $t, t' \in I$,

$$\begin{aligned}
&|\mathcal{N}^\ell_{x,I,q,\delta}(t,\bar{z}_x;y) - \mathcal{N}^\ell_{x',I,q,\delta}(t',\bar{z}_{x'};y')| \le (A + |\underline{\omega}|_1\delta)|t-t'| \\
&\quad + \max\big\{|y-y'|, |x-x'|\big\} e^{L_\delta|I|}.
\end{aligned} \tag{6.59}$$

Proof : Recall from (6.52) that $\mathcal{N}_{x,I,q,\delta}(t,\bar{z};y) := x + \sum_{i=1}^{q}\rho_i(t)\sum_{j=1}^{d_y} y_j\omega_j \mathcal{N}_{j,i,\delta}(\bar{z}(\xi_i))$, where $\mathcal{N}_{j,i,\delta}$ are Lipschitz stable DNNs approximating $\mathbf{a}^{\circ}_{j,J_i}(\cdot)$ (Proposition 2.14). Thus, by (3.7),

$$\begin{aligned}\sum_{j=1}^{d_y}\omega_j|\mathcal{N}_{j,i,\delta}(\bar{z}(\xi_i))| &\le \sum_{j=1}^{d_y}|\mathbf{a}_{j,J_i}(\bar{z}(\xi_i))| + \omega_j|\mathbf{a}^{\circ}_{j,J_i}(\bar{z}(\xi_i)) - \mathcal{N}_{j,i,\delta}(\bar{z}(\xi_i))|\\ &\le A + |\underline{\omega}|\delta, \end{aligned} \tag{6.60}$$

while, by Proposition 2.14, (2.19),

$$|\mathcal{N}_{j,i,\delta}(\bar{z}) - \mathcal{N}_{j,i,\delta}(\bar{z}')| \le \bar{\Lambda}\|\bar{z}-\bar{z}'\|_{L_\infty(I;\mathbb{R}^m)}, \quad \bar{z},\bar{z}'\in L_\infty(I;\mathbb{R}^m),$$

uniformly in $i = 1,\dots,d_y,\ \ j = 1,\dots,d_y$, with $\bar{\Lambda}$ from (6.57). Then

$$\begin{aligned}&\left|\mathcal{N}_{x,I,q,\delta}(t,\bar{z};y) - \mathcal{N}_{x,I,q,\delta}(t,\bar{z}';y')\right|\\ &\quad\le \sum_{i=1}^{q}\rho_i(t)\sum_{j=1}^{d_y}\omega_j\left|y_j\mathcal{N}_{j,i,\delta}(\bar{z}(\xi_i)) - y'_j\mathcal{N}_{j,i,\delta}(\bar{z}'(\xi_i))\right|\\ &\quad\le \sum_{i=1}^{q}\rho_i(t)\Big\{\sum_{j=1}^{d_y}\omega_j|y_j-y'_j||\mathcal{N}_{j,i,\delta}(\bar{z}(\xi_i))|\\ &\qquad + \sum_{j=1}^{d_y}\omega_j|y'_j||\mathcal{N}_{j,i,\delta}(\bar{z}(\xi_i)) - \mathcal{N}_{j,i,\delta}(\bar{z}'(\xi_i))|\Big\}\\ &\quad\le \sum_{i=1}^{q}\rho_i(t)\Big\{(A+|\underline{\omega}|_1\delta)|y-y'| + \bar{\Lambda}|\underline{\omega}|_1\|\bar{z}-\bar{z}'\|_{L_\infty(I;\mathbb{R}^m)}\Big\}\\ &\quad\le \sum_{i=1}^{q}\rho_i(t)L_\delta\max\{|y-y'|,\|\bar{z}-\bar{z}'\|_{L_\infty(I;\mathbb{R}^m)}\}.\end{aligned}$$

Similarly, for $t,t'\in I$

$$\begin{aligned}|\mathcal{N}_{x,I,q,\delta}(t,\bar{z};y) - \mathcal{N}_{x,I,q,\delta}(t',\bar{z};y)| &\le \sum_{i=1}^{q}|\rho_i(t)-\rho_i(t')|\sum_{j=1}^{d_y}|y_j|\omega_j|\mathcal{N}_{j,i,\delta}(\bar{z}(\xi_i))|\\ &\le (A+|\underline{\omega}|_1\delta)|t-t'|,\end{aligned}$$

where we have used (6.20) and (6.60). Applying this to $\bar{z} = \mathcal{N}^{\ell-1}_{\cdot,I,q,\delta}$ extends this to iterates of $\mathcal{N}_{\cdot,I,q,\delta}$.

Hence, we are in the same situation as in (6.77). Therefore, (6.58) and (6.59) follow by the same arguments as used in the proof of Lemma 6.6. □

Now recall that $\delta(\varepsilon) \eqsim \frac{\varepsilon e^{-K/2}}{2|I||\underline{\omega}|_1}$. Hence, $\delta(\varepsilon)|\underline{\omega}|_1 \lesssim \frac{\varepsilon e^{-K/2}}{2|I|} \lesssim \widehat{T}^{-1}\varepsilon K e^{-K/2} \leq \varepsilon/\widehat{T}$. Then

$$L_{\delta(\varepsilon)} \leq A + \bar{\Lambda}|\underline{\omega}|_1 + \widehat{T}^{-1}\varepsilon \leq \widehat{L} := \widehat{A} + \bar{\Lambda}|\underline{\omega}|_1, \quad \widehat{A} := \max_{\varepsilon\leq 1} A + \widehat{T}^{-1}\varepsilon,$$

and the same arguments as used earlier provide

$$|\widehat{\Psi}_{\underline{\tau}^k(\varepsilon)}(t,x,y) - \widehat{\Psi}_{\underline{\tau}^k(\varepsilon)}(t',x',y')|_\infty \leq \widehat{A}|t-t'| + \widehat{L}\max\{|x-x'|,|y-y'|\}.$$

From these observations it follows that $|||\mathcal{N}_\varepsilon|||_{N_\varepsilon,m} \leq e^{\widehat{L}\widehat{T}}$ and for $\gamma(r) := \left(\frac{r}{d_y}\right)^{\frac{1}{m+1}}\left|\log_2 \frac{r}{d_y}\right|^{-\frac{2}{m+1}}$

$$\gamma(N_\varepsilon)K_m(z,N,\gamma(N)^{-1}) \leq \gamma(N_\varepsilon)\|z-\mathcal{N}_\varepsilon\|_{L_\infty(\Omega)} + |||\mathcal{N}_\varepsilon|||_{N_\varepsilon,m} \lesssim e^{LT} + e^{\widehat{L}\widehat{T}},$$

which confirms the remainder of the assertion. □

6.6 Proof of Theorem 4.9

By assumption, given $\tilde{\varepsilon} > 0$, there exists an $f_{\tilde{\varepsilon}} \in L_\infty(\widehat{I}; \mathfrak{C}_{N_{\tilde{\varepsilon}}(f),m}) \cap \mathrm{Lip}_1(\widehat{I}; C(\mathbb{R}^m \times \mathcal{Y}))$, piecewise affine in time, and a composition $u_{0,\tilde{\varepsilon}} \in \mathfrak{C}_{N_{\tilde{\varepsilon}}(u_0),m}$, so that (identifying for notational convenience in what follows mappings and representations)

$$\begin{aligned} &\|u_0 - u_{0,\tilde{\varepsilon}}\|_{L_\infty(\mathbb{R}^m\times\mathcal{Y})} \leq \tilde{\varepsilon}, \\ &\mathfrak{N}(u_{0,\tilde{\varepsilon}}) \leq \gamma^{-1}\big(\|u_0\|_{\mathcal{A}^{\gamma,m}}/\tilde{\varepsilon}\big) \eqsim \|u_0\|^{1/\alpha}_{\mathcal{A}^{\gamma,m}}\tilde{\varepsilon}^{-1/\alpha}, \end{aligned} \tag{6.61}$$

and likewise, since $\mathfrak{N}(f_{\tilde{\varepsilon}}(t,\cdot)) \leq \gamma^{-1}\big(\|f(t,\cdot)\|_{\mathcal{A}^{\gamma,m}}/\tilde{\varepsilon}\big)$, for $t \in \widehat{I}$

$$\|f - f_{\tilde{\varepsilon}}\|_{L_\infty(\widehat{I};C(\mathbb{R}^m\times\mathcal{Y}))} \leq \tilde{\varepsilon}, \quad \mathfrak{N}(f_{\tilde{\varepsilon}}(t,\cdot)) \lesssim \|f(t,\cdot)\|^{1/\alpha}_{\mathcal{A}^{\gamma,m}}\tilde{\varepsilon}^{-\frac{1}{\alpha}}. \tag{6.62}$$

Next, we use that the compositional factors in $u_{0,\tilde{\varepsilon}}$, $f_{\tilde{\varepsilon}}$ are Lipschitz continuous with constants controlled by $\|u_0\| := \|u_0\|_{\mathcal{A}^{\gamma,m}}$, $\|f\| := \|f\|_{L_\infty(\widehat{I};\mathcal{A}^{\gamma,m})}$, respectively. We employ Lemma 2.1 to implant η-accurate Lipschitz controlled DNNs into $u_{0,\tilde{\varepsilon}}$, $f_{\tilde{\varepsilon}}$, respectively. Invoking Remark 2.13, we obtain

$$\|u_{0,\tilde{\varepsilon}} - \mathcal{N}_{u_{0,\tilde{\varepsilon}},\eta}\|_{L_\infty(\mathbb{R}^m\times\mathcal{Y})} \leq \eta\Big\{1 + \sum_{j=1}^{n(\mathsf{D}(u_{0,\tilde{\varepsilon}}))-1}\|u_0\|\Big\} \leq \eta\mathfrak{N}(u_{0,\tilde{\varepsilon}})\|u_0\|, \tag{6.63}$$

as well as

$$\begin{aligned} \|f_{\tilde{\varepsilon}}(t,\cdot) - \mathcal{N}_{f_{\tilde{\varepsilon}},\eta}(t,\cdot)\|_{L_\infty(\mathbb{R}^m\times\mathcal{Y})} &\leq \eta\Big\{1 + \sum_{j=1}^{n(\mathsf{D}(f_{\tilde{\varepsilon}}))-1}\|f_0\|\Big\} \\ &\leq \eta\mathfrak{N}(f_{\tilde{\varepsilon}}(t))\|f_0\|, \quad t \in I. \end{aligned} \tag{6.64}$$

By Proposition 2.14, (6.61), and (6.62), it follows that

$$\begin{aligned}\#\mathcal{N}_{u_{0,\tilde{\varepsilon}},\eta} &\le \|u_{0,\tilde{\varepsilon}}\|^m_{\mathfrak{N}(u_{0,\tilde{\varepsilon}}),m}\,\mathfrak{N}(u_{0,\tilde{\varepsilon}})\eta^{-m}|\log_2\eta| \\ &\le \|u_0\|^{m+\frac{1}{\alpha}}\tilde{\varepsilon}^{-\frac{1}{\alpha}}\eta^{-m}|\log_2\eta|, \end{aligned} \tag{6.65}$$

and, uniformly in $t\in\widehat{I}$,

$$\begin{aligned}\#\mathcal{N}_{f_{\tilde{\varepsilon}},\eta}(t,\cdot) &\le \|f_{\tilde{\varepsilon}}(t,\cdot)\|^m_{\mathfrak{N}(f_{\tilde{\varepsilon}}(t)),m,\mathrm{Lip}}\,\mathfrak{N}(f_{\tilde{\varepsilon}}(t))\eta^{-m}|\log_2\eta| \\ &\le \|f\|^{m+\frac{1}{\alpha}}\tilde{\varepsilon}^{-\frac{1}{\alpha}}\eta^{-m}|\log_2\eta|. \end{aligned} \tag{6.66}$$

Employing again a time-discretization of size q, (3.19) suggests the following DNN approximation to u which yields, on account of Lemma 6.3,

$$\Big|\int_0^t f_{\tilde{\varepsilon}}(s,w,y))ds - \sum_{i=1}^q \rho_i(t)\,f_{\tilde{\varepsilon}}(\xi_i,w,y)\Big| \le \frac{\|f\|\widehat{T}^2}{2q}. \tag{6.67}$$

Finally, we know from Theorem 4.7 that there exists a DNN $\mathcal{N}_{z,\tilde{\varepsilon}_z}$ that approximates the characteristic field z within accuracy $\tilde{\varepsilon}$, i.e., in view of (4.8),

$$\|z-\mathcal{N}_{z,\tilde{\varepsilon}}\|_{L_\infty(\Omega)}\le\tilde{\varepsilon},\quad \#\mathcal{N}_{z,\tilde{\varepsilon}} \lesssim d_y\widehat{T}\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)^{m+1}\Big|\log_2\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)\Big|^2,$$

where we suppress in what follows the dependence on A, m and where L is given by (4.7).

In summary, the network $\mathcal{N}_{u,\tilde{\varepsilon},\eta,q}$ formed by composing the DNNs $\mathcal{N}_{u_0,\tilde{\varepsilon},\eta}$, $\mathcal{N}_{f_{\tilde{\varepsilon}},\eta}(\xi_i,\cdot)$, $i=1,\dots,q$, with the approximate characteristics $\mathcal{N}_{z,\tilde{\varepsilon}_z}$ satisfies, on account of (2.4), (2.3), (6.65), and (6.66),

$$\begin{aligned}\#\mathcal{N}_{u,\tilde{\varepsilon},\eta,q} &\le \#\big(\mathcal{N}_{u_0,\tilde{\varepsilon},\eta}\circ\mathcal{N}_{z,\tilde{\varepsilon}}\big) + q\max_{i=1,\dots,q}\#\big(\mathcal{N}_{f_{\tilde{\varepsilon}},\eta}(\xi_i,\cdot)\circ\mathcal{N}_{z,\tilde{\varepsilon}}\big) \\ &= \#\mathcal{N}_{u_0,\tilde{\varepsilon},\eta} + \#\mathcal{N}_{z,\tilde{\varepsilon}} + q\big(\max_{i=1,\dots,q}\#\mathcal{N}_{f_{\tilde{\varepsilon}}\,\eta}(\xi_i,\cdot) + \#\mathcal{N}_{z,\tilde{\varepsilon}}\big) \\ &\lesssim \Big(\|u_0\|^{m+\frac{1}{\alpha}}\tilde{\varepsilon}^{-\frac{1}{\alpha}} + q\|f\|^{m+\frac{1}{\alpha}}\tilde{\varepsilon}^{-\frac{1}{\alpha}}\Big)\eta^{-m}|\log_2\eta| \\ &\quad + (1+q)d_y\widehat{T}\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)^{m+1}\Big|\log_2\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)\Big|^2 \\ &\lesssim q\Big\{M^{m+\frac{1}{\alpha}}\tilde{\varepsilon}^{-\frac{1}{\alpha}}\eta^{-m}|\log_2\eta| \\ &\quad + d_y\widehat{T}\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)^{m+1}\Big|\log_2\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)\Big|^2\Big\}, \end{aligned} \tag{6.68}$$

where we set $M:=\max\{1,\|u_0\|,\|f\|\}$. To determine η, we have by (3.19),

$$\begin{aligned}\|u(t,\cdot)-\mathcal{N}_{u,\tilde{\varepsilon},\eta,q}(t,\cdot)\|_{L_\infty(\mathbb{R}^m\times\mathcal{Y})} &\le \|u_0(z(-t,t,\cdot;\cdot))\\ &\quad -\mathcal{N}_{u_0,\tilde{\varepsilon},\eta}\circ\mathcal{N}_{z,\tilde{\varepsilon}}(-t,t,\cdot;\cdot)\|_{L_\infty(\mathbb{R}^m\times\mathcal{Y})}\\ &\quad +\sup_{x,y}\left|\int_0^t f(s,z(t-s,t,x;y);y)ds-\sum_{i=1}^q\rho_i(t)\mathcal{N}_{\xi_i,f_{\tilde{\varepsilon}},\eta}\circ\mathcal{N}_{z,\tilde{\varepsilon}}(t-\xi_i,t,x;y)\right|\\ &=: Q_1+\sup_{x,y}Q_2(x,y).\end{aligned}$$

Regarding Q_1, let $L_0 := |u_0|_{\mathrm{Lip}_1(\mathbb{R}^m\times\mathcal{Y})} \le \|u_0\|$. Because of (6.63) and (6.61),

$$\begin{aligned}Q_1 &\le \|u_0(z(-t,t,\cdot;\cdot))-u_0(\mathcal{N}_{z,\tilde{\varepsilon}}(-t,t,\cdot;\cdot))\|_{L_\infty(\mathbb{R}^m\times\mathcal{Y})}\\ &\quad +\|u_0(\mathcal{N}_{z,\tilde{\varepsilon}}(-t,t,\cdot;\cdot))-u_{\tilde{\varepsilon},0}(\mathcal{N}_{z,\tilde{\varepsilon}}(-t,t,\cdot;\cdot))\|_{L_\infty(\mathbb{R}^m\times\mathcal{Y})}\\ &\quad +\|u_{\tilde{\varepsilon},0}(\mathcal{N}_{z,\tilde{\varepsilon}}(-t,t,\cdot;\cdot))-\mathcal{N}_{u_0,\tilde{\varepsilon},\eta}(\mathcal{N}_{z,\tilde{\varepsilon}}(-t,t,\cdot;\cdot))\|_{L_\infty(\mathbb{R}^m\times\mathcal{Y})}\\ &\le (1+L_0)\tilde{\varepsilon}+\|u_0\|\eta\mathfrak{N}(u_{0,\tilde{\varepsilon}}) \;\lesssim\; (1+L_0)\tilde{\varepsilon}+\|u_0\|^{1+\frac{1}{\alpha}}\eta\tilde{\varepsilon}^{-\frac{1}{\alpha}}\\ &\le \|u_0\|\Big\{2\tilde{\varepsilon}+\eta\tilde{\varepsilon}^{-\frac{1}{\alpha}}\|u_0\|^{\frac{1}{\alpha}}\Big\},\end{aligned}\tag{6.69}$$

where we have used $\|u_0\| \ge 1$ and (4.10) in the last step. Similarly, by (6.62) (6.64), and (6.67), abbreviating $L_f := |f|_{\mathrm{Lip}_1(\widehat{T};C(\mathbb{R}^m\times\mathcal{Y}))}$,

$$\begin{aligned}Q_2(x,y) &\le \widehat{T}L_f\tilde{\varepsilon}+\int_0^t|f(s,\mathcal{N}_{z,\tilde{\varepsilon}}(t-s,t,x,y));y-f_{\tilde{\varepsilon}}(s,\mathcal{N}_{z,\tilde{\varepsilon}}(t-s,t,x,y);y)|ds\\ &\quad +\int_0^t\Big|f_{\tilde{\varepsilon}}(s,\mathcal{N}_{z,\tilde{\varepsilon}}(t-s,t,x,y);y)-\sum_{i=1}^q\rho_i(t)f_{\tilde{\varepsilon}}(\xi_i,\mathcal{N}_{z,\tilde{\varepsilon}}(t-\xi_i,x,y);y)\Big|\\ &\quad +\sum_{i=1}^q\rho_i(t)\big|f_{\tilde{\varepsilon}}(\xi_i,\mathcal{N}_{z,\tilde{\varepsilon}}(t-\xi_i,x,y);y)-\mathcal{N}_{f_{\tilde{\varepsilon}},\eta}(\xi_i,\mathcal{N}_{z,\tilde{\varepsilon}}(t-\xi_i,x,y);y)\big|.\end{aligned}$$

On account of (6.22) and the assumption $L_f \le \|f\|$, this gives

$$\begin{aligned}Q_2(x,y) &\le (1+\|f\|)\widehat{T}\tilde{\varepsilon}+\frac{\|f\|\widehat{T}^2}{2q}+\widehat{T}\eta\mathfrak{N}(f_{\tilde{\varepsilon}})\|f\|\\ &\le 2\|f\|\widehat{T}\tilde{\varepsilon}+\frac{\|f\|\widehat{T}^2}{2q}+\widehat{T}\frac{\eta}{\tilde{\varepsilon}^{1/\alpha}}\|f\|^{1+\frac{1}{\alpha}}\\ &\le \widehat{T}\|f\|\Big\{2\tilde{\varepsilon}+\frac{\widehat{T}}{2q}+\|f\|^{\frac{1}{\alpha}}\eta\tilde{\varepsilon}^{-\frac{1}{\alpha}}\Big\}.\end{aligned}\tag{6.70}$$

Now recall that $M = \max\{1, \|f\|, \|u_0\|\}$ and let

$$q(\tilde{\varepsilon}) = \frac{\widehat{T}}{2\tilde{\varepsilon}}, \quad \eta(\tilde{\varepsilon}) = M^{-\frac{1}{\alpha}}\tilde{\varepsilon}^{1+\frac{1}{\alpha}},\tag{6.71}$$

to conclude that $\max_{x,y} Q_2(x,y) \le 4\widehat{T}\|f\|\tilde{\varepsilon}$. Hence, we derive from (6.69) and (6.70) that the network $\mathcal{N}_{u,\tilde{\varepsilon}} := \mathcal{N}_{u,\tilde{\varepsilon},\eta(\tilde{\varepsilon}),q(\tilde{\varepsilon})}$ satisfies (recall that by assumptions $\|u_0\|, \|f\| \ge 1$)

$$\|u - \mathcal{N}_{u,\tilde{\varepsilon}}\|_{L_\infty(\Omega)} \le \{4\widehat{T}\|f\| + 3\|u_0\|\}\tilde{\varepsilon} \le 7\widehat{T}M\tilde{\varepsilon}.$$

This confirms the first part of (4.12) with $\varepsilon := 7\widehat{T}M\tilde{\varepsilon}$.

Now we infer from (6.68), (6.61), (6.62) that

$$\begin{aligned}
\#\mathcal{N}_{u,\tilde{\varepsilon}} &\lesssim \widehat{T}\tilde{\varepsilon}^{-1}\Big\{M^{\frac{\alpha m+1}{\alpha}}\tilde{\varepsilon}^{-\frac{1}{\alpha}}M^{\frac{m}{\alpha}}\tilde{\varepsilon}^{-\frac{m(1+\alpha)}{\alpha}}\Big|\log_2\frac{M}{\tilde{\varepsilon}^{\alpha+1}}\Big| \\
&\quad + d_y\widehat{T}\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)^{m+1}\Big|\log_2\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)\Big|^2\Big\} \\
&\lesssim \widehat{T}\tilde{\varepsilon}^{-1}\Big\{M^{\frac{\alpha m+1}{\alpha}}\tilde{\varepsilon}^{-\frac{1}{\alpha}}M^{\frac{m}{\alpha}}\tilde{\varepsilon}^{-\frac{m(1+\alpha)}{\alpha}} + d_y\widehat{T}e^{\widehat{T}L(m+1)}\tilde{\varepsilon}^{-(m+1)}\Big\}\Big|\log_2\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)\Big|^2 \\
&= \widehat{T}\Big\{M^{\frac{(\alpha+1)m+1}{\alpha}}\tilde{\varepsilon}^{-\frac{(1+\alpha)(m+1)}{\alpha}} + d_y\widehat{T}e^{\widehat{T}L(m+1)}\tilde{\varepsilon}^{-(m+2)}\Big\}\Big|\log_2\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)\Big|^2.
\end{aligned}$$

Introducing $\beta := \max\{1, (m+1)/\alpha\}$, (see (4.11)) and substituting $\tilde{\varepsilon} = \varepsilon/(7\widehat{T}M)$, yields upon elementary calculations

$$\begin{aligned}
\#\mathcal{N}_{u,\tilde{\varepsilon}} &\lesssim \Big\{M^{\frac{(\alpha+1)m+1}{\alpha}}\widehat{T}^{m+2+\beta}e^{-L\widehat{T}(m+1+\beta)} \\
&\quad + d_y\widehat{T}^{m+4}e^{-L\widehat{T}\beta}\Big\}\Big(\frac{Me^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)^{m+1+\beta}\Big|\log_2\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)\Big|^2.
\end{aligned}$$

The terms $\widehat{T}^{m+2+\beta}e^{-L\widehat{T}(m+1+\beta)}$, $\widehat{T}^{m+4}e^{-L\widehat{T}\beta}$ remain uniformly bounded for all $\widehat{T} > 0$ with a constant that actually decreases when L gets large. Thus, fixing M, a large parametric dimension in the second summand dominates, giving

$$\begin{aligned}
\#\mathcal{N}_{u,\tilde{\varepsilon}} &\lesssim \max\{M^{\frac{(\alpha+1)m+1}{\alpha}}, d_y\}\Big(\frac{Me^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)^{m+1+\beta}\Big|\log_2\Big(\frac{e^{L\widehat{T}}}{\tilde{\varepsilon}}\Big)\Big|^2 \\
&=: \phi(Me^{L\widehat{T}}/\varepsilon),
\end{aligned} \tag{6.72}$$

which proves (4.12).

Regarding the remainder of the claim, recall from Theorem 4.7 that the approximations $\mathcal{N}_{z,\varepsilon}$ have uniformly bounded composition norms $|||\mathcal{N}_{z,\varepsilon}|||_{\#\mathcal{N}_{z,\varepsilon},m} \lesssim e^{L\widehat{T}}$, see also Lemma 6.8. To bound the composition norms of $\mathcal{N}_{u,\tilde{\varepsilon}}$, we recall that the composition norms of the network approximations to u_0 and f are bounded by $M = \max\{1, \|u_0\|, \|f\|\}$. We then infer from Remark 2.12 (see also (2.10) and (2.9)), applied to the first line of (6.68), that

$$|||\mathcal{N}_{u,\tilde{\varepsilon}}|||_{\#\mathcal{N}_{u,\tilde{\varepsilon}},m} \lesssim Me^{L\widehat{T}},$$

which is the asserted stability estimate. The convergence rate (4.13) follows from Remark 6.1 and Lemma 6.1 applied to to the growth function ϕ in (6.72). □

Regarding Remark 4.10, the same reasoning applies (with slightly simpler technicalities), replacing (6.71) by $\eta(\tilde{\varepsilon}) := \tilde{\varepsilon}/(\ln((\|u_0\| + \|f\|)/\tilde{\varepsilon}))$ while keeping $q(\tilde{\varepsilon})$ the same, □

Appendix A

Proof of Proposition 2.14 In this section we build mainly on findings from [19, 37]. Consider the "hat-function" $\phi(x) := (1 - |x|)_+ = \max\{0, 1 - |x|\}$, $x \in \mathbb{R}$, as well as the scaled and shifted versions $\phi_{i,h}(x) := \phi(h^{-1}x - i)$, $i \in \mathbb{Z}$, with support $S_i = [(i-1)h, (i+1)h]$. We let $h = 1/q$ for some integer $q \in \mathbb{N}$, so that the restrictions of the $\phi_{i,h}$ to $[0, 1]$ form a stable basis for all piecewise linears on $(0, 1)$ subordinate to the partition induced by the nodes $\{ih = i/q : i = 0, \dots, q\}$. Since each $\phi_{i,h}$ is a second order divided difference of the ReLU rectifier $\sigma(x) := x_+$ with respect to the nodes ih, it has an exact representation as a univariate neural network of fixed finite depth and a fixed finite number of weights. The Lipshitz constant of $\phi_{i,h}$ and hence of this network is clearly h^{-1}. Abbreviating $\mathbf{i} := (i_1, \dots, i_s) \in \{0, \dots, q\}^s$ we consider next for $x = (x_1, \dots, x_s) \in \mathbb{R}^s$ the tensor products $\phi_{\mathbf{i},h} := \phi_{i_1,h}(x_1) \cdots \phi_{i_s,h}(x_s)$, which obviously satisfy $\phi_{\mathbf{i},h}(\mathbf{i}') = \delta_{\mathbf{i},\mathbf{i}'}$ for any $\mathbf{i}, \mathbf{i}' \in \mathcal{I}_h := \{0, h, \dots, hq\}^s$ while we still have $\|\partial_j \phi_{\mathbf{i},h}\|_\infty \le h^{-1}$. The next step consists in approximating each $\phi_{\mathbf{i},h}$, viz. a product of univariate ReLU networks of fixed depth and number of weights by a ReLU network of input dimension s. This is where one uses that the function $M : v = (v_1, \dots, v_s) \mapsto \prod_{j=1}^s v_s$ can be approximated by a ReLU network $\mathcal{N}_{M,\delta}$ according to

$$\|M - \mathcal{N}_{M,\delta}\|_{W^k(L_\infty((0,1)^s))} \le \delta, \quad k \in \{0, 1\}, \tag{6.73}$$

where the depth of $\mathcal{N}_{M,\delta}$ as well as $\#\mathcal{N}_{M,\delta}$ is bounded by a constant multiple of $\log_2 \delta^{-1}$, with constants depending only on s. Moreover, $\mathcal{N}_{M,\delta}(0) = 0$. The case $k = 0$ in (6.73) appears already in [37]. A key observation in [19, § C] is that $k = 1$ still holds under the same complexity bounds. This is then used to show that for each $\mathbf{i} \in \mathcal{I}_h$ there exists a ReLU network $\mathcal{N}_{\mathbf{i},\delta}$ such that

$$\begin{gathered} \|\phi_{\mathbf{i},h} - \mathcal{N}_{\mathbf{i},\delta}\|_{W^k(L_\infty((0,1)^s))} \le c^k \delta h^{-1}, \quad k \in \{0, 1\}, \\ \|\mathcal{N}_{\mathbf{i},\delta}\|_{\mathrm{Lip}_1} \le c\delta^{-1}, \quad \operatorname{supp} \mathcal{N}_{\mathbf{i},\delta} \subseteq \operatorname{supp} \phi_{\mathbf{i},h}, \end{gathered} \tag{6.74}$$

(with a constant c, depending on s) and

$$\#\mathcal{N}_{\mathbf{i},\delta}, \ \text{depth of } \mathcal{N}_{\mathbf{i},\delta} \ \lesssim \ \log_2 \frac{1}{\delta},$$

with constants depending only on s. Now, given v, consider the interpolant

$$v_h := \sum_{\mathbf{i} \in \mathcal{I}_h} v(\mathbf{i}h)\phi_{\mathbf{i},h}.$$

Obviously $\|v_h\|_\infty \le \|v\|_\infty$. We claim that v_h is also Lipschitz continuous. To see this, let $\mathcal{I}_h(x) := \{\mathbf{i} \in \mathcal{I}_h : \phi_{\mathbf{i},h}(x) \neq 0\}$ denote the collection of those nodes whose basis functions contain $x \in (0,1)^{\mathrm{s}}$ in the interior of their support. Then for $x \in (0,1)^{\mathrm{s}}$ let $\mathbf{i}(x) \in \mathcal{I}_h(x)$ denote the node closest to the Chebyshev center of the convex hull $\big[\mathcal{I}_h(x)\big]$ of $\mathcal{I}_h(x)$. Then, since $v(\mathbf{i}(x)h)\sum_{\mathbf{i}\in\mathcal{I}_c(x)}\phi_{\mathbf{i},h}(x)$ is constant in a neighborhood of x, one has

$$\begin{aligned}|\partial_j v_h(x)| &= \Big|\sum_{\mathbf{i}\in\mathcal{I}_h(x)} (v(\mathbf{i}h) - v(\mathbf{i}(x)h))\partial_j\phi_{\mathbf{i},h}(x)\Big| \\ &\le \max_{\mathbf{i}\in\mathcal{I}_h(x)} |v(\mathbf{i}h) - v(\mathbf{i}(x)h)|\frac{1}{h}\#(\mathcal{I}_h(x)).\end{aligned}$$

Since $\operatorname{diam}\mathcal{I}_h(x) \le ch$ for a constant depending on s, this yields

$$|\partial_j v_h(x)| \le c^{-1}\|v\|_{\mathrm{Lip}_1}, \quad x \in (0,1)^{\mathrm{s}}, \tag{6.75}$$

from which it follows that (weakly)

$$\begin{aligned}|v_h(x) - v_h(x')| =& \Big|\int_0^1 \nabla v_h(x + t(x'-x))\cdot(x'-x)dt\Big| \\ \le& |x-x'|\sqrt{\mathrm{s}}c^{-1}\|v\|_{\mathrm{Lip}_1}.\end{aligned}$$

Moreover,

$$|v(x) - v_h(x)| = \Big|\sum_{i\in\mathcal{I}_h(x)} (v(x) - v(\mathbf{i}h))\phi_{\mathbf{i},h}\Big| \le Ch\|v\|_{\mathrm{Lip}_1},$$

since $\max\{|x - \mathbf{i}h| : \mathbf{i} \in \mathcal{I}_h(x)\} \le Ch$ whith C depending only on s. Given $\delta > 0$, the choice $h = h(\delta) \le \frac{\delta}{2C\|v\|_{\mathrm{Lip}_1}}$ ensures $|v(x) - v_h(x)| \le \frac{\delta}{2}$. Now approximate each $\phi_{\mathbf{i},h(\delta)}$ by a ReLU network $\mathcal{N}_{\mathbf{i},\delta}$ with accuracy $\|\phi_{\mathbf{i},h(\delta)} - \mathcal{N}_{\mathbf{i},\delta}\|_\infty \le c^*\delta$, with c^* to be determined in a moment. We obtain, by (6.74)

$$\begin{aligned}\Big\|v - \sum_{\mathbf{i}\in\mathcal{I}_{h(\delta)}} v(\mathbf{i}h)\mathcal{N}_{\mathbf{i},\delta}\|_\infty \le& \|v - v_h\|_\infty + \sup_{x\in(0,1)^{\mathrm{s}}}\sum_{\mathbf{i}\in\mathcal{I}_{h(\delta)}(x)} |v(\mathbf{i}h)|\big|\phi_{\mathbf{i},h(\delta)}(x) - \mathcal{N}_{\mathbf{i},\delta}(x)\big| \\ \le& \frac{\delta}{2} + \sup_{x\in(0,1)^{\mathrm{s}}} \#\mathcal{I}_{h(\delta)}(x)c^*\delta.\end{aligned}$$

Thus, choosing $c^* = (2\sup_{x\in(0,1)^{\mathrm{s}}}\#\mathcal{I}_{h(\delta)}(x))^{-1}$, we have confirmed

$$\Big\|v - \sum_{\mathbf{i}\in\mathcal{I}_{h(\delta)}} v(\mathbf{i}h)\mathcal{N}_{\mathbf{i},\delta}\|_\infty \le \delta.$$

Moreover, defining $\mathcal{N}_\delta := \sum_{\mathbf{i}\in\mathcal{I}_h} v(\mathbf{i}h)\mathcal{N}_{\mathbf{i},\delta}$, we obtain, again by (6.74) and (6.75),

$$\begin{aligned}\left|\partial_j \mathcal{N}_\delta(x)\right| &\le \left|\partial_j(\mathcal{N}_\delta(x) - v_h(x))\right| + c^{-1}\|v\|_{\mathrm{Lip}_1}\\ &= \Big|\sum_{\mathbf{i}\in\mathcal{I}_{h(\delta)}} v(\mathbf{i}h)\big(\partial_j\mathcal{N}_{\mathbf{i},\delta}(x) - \partial_j\phi_{\mathbf{i},h}\big)\Big| + c^{-1}\|v\|_{\mathrm{Lip}_1}\\ &\le c'\|v\|_\infty \#(\mathcal{I}_{h(\delta)})h^{-1}\delta + c^{-1}\|v\|_{\mathrm{Lip}_1} \le c''(\|v\|_\infty + 1)\|v\|_{\mathrm{Lip}_1},\end{aligned}$$

where c'' depends only on s and where we have used that $\frac{\delta}{h} \le 4C\|v\|_{\mathrm{Lip}_1}$. This confirms (2.19). Regarding the complexity (2.20) of $\mathcal{N}_\delta$, we have $\#(\mathcal{I}_h(\delta)) = h(\delta)^{-\mathrm{s}} \le c\delta^{-\mathrm{s}}\|v\|^{\mathrm{s}}_{\mathrm{Lip}_1}$, which completes the proof because $\#\mathcal{N}_\delta \lesssim \#(\mathcal{I}_h(\delta))\log_2\frac{1}{\delta}$. □

Proof of Proposition 2.9 For each finite N there is only a finite number of feasible dimensionality vectors D with $\mathfrak{N}(\mathsf{D}) \le N$ which representations of $G \in \mathfrak{C}_{N,\mathrm{s}}$ may have. For each such D consider first the following auxiliary classes. Let $\mathbb{F}_\ell \subset C(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})$ be compact and let $\mathfrak{C}(\mathsf{D},\mathbb{F}) := \{G \in \mathbb{X}_0, G = G_{\mathbf{g}}, \mathsf{D}(\mathbf{g}) = \mathsf{D} : \text{with } g^j \in \mathbb{F}_j,\ j = 1,\dots,n(\mathsf{D})\}$. □

Lemma 6.9 *The collection*

$$\mathfrak{C}_{N,\mathrm{s}}(\mathbb{F}) := \bigcup_{\mathfrak{N}(D)\le N} \mathfrak{C}(D,\mathbb{F})$$

is compact in $C(D_0;\mathbb{R}^{d_{n(D)}})$ *(and so are the subsets* $\mathfrak{C}(D,\mathbb{F})$*).*

To establish first this Lemma, it is enough to confirm compactness of $\mathfrak{C}(\mathsf{D},\mathbb{F})$ for each of the eligible D. To this end, let $\mathbb{F}$, $\mathbb{F}'$, be dimensionally compatible compact subclasses so that for $g \in \mathbb{F}$, $h \in \mathbb{F}'$ compositions $h \circ g$ are defined. Then $\{h \circ g : h \in \mathbb{F}',\ g \in \mathbb{F}\}$ is compact in $\mathbb{X}$. In fact, let $(g_n)_{n\in\mathbb{N}}$, $(h_n)_{n\in\mathbb{N}}$ be uniformly bounded sequences in $\mathbb{F}$, $\mathbb{F}'$, respectively. By Arzela-Ascoli's Theorem, they are equicontinuous and have a convergent subsequence with continuous limits g, h, say. Denote these subsequences again by $(g_n)_{n\in\mathbb{N}}$, $(h_n)_{n\in\mathbb{N}}$). Then,

$$\|h_n \circ g_n - h \circ g\|_{\mathbb{X}(D_g)} \le \|h_n \circ g_n - h_n \circ g\|_{\mathbb{X}(D_g)} + \|h_n \circ g - h \circ g\|_{\mathbb{X}(D_g)}.$$

By uniform convergence of the g_n and equicontinuity of the h_n the first summand becomes arbitrarily small for n large enough. By uniform convergence of the h_n the second summand gets small as well. Iterating this argument, shows that $\mathfrak{C}(\mathsf{D},\mathbb{F})$ is compact. Since $\mathfrak{C}_{N,\mathrm{s}}(\mathbb{F})$ is a finite union of such sets the assertion of Lemma 6.9 follows.

The proof of Proposition 2.9 follows now from noticing that membership to $\mathfrak{C}_{N,\mathrm{s}}(B)$ requires all composition factors to have uniformly bounded Lipschitz norm and hence belong to compact classes. □

Proof of Remark 2.10: Let $(f_j)_{j\in\mathbb{N}}$ be a sequence in $U\mathcal{A}^{\gamma,\mathrm{s}}$. Take a sequence $(\varepsilon_k)_{k\in\mathbb{N}}$ of numbers decreasing monotonically to zero. For each f_j let $\mathbf{g}_{j,k}$ denote a compositional representation of some function in $\mathfrak{C}_{N_{\varepsilon_k},\mathrm{s}}$ such that (see (2.12)) $\|\!|G_{\mathbf{g}_{j,k}}|\!\|_{N_{\varepsilon_k},\mathrm{s},\mathcal{R}^\circ} \le 1$

and $\|f_j - G_{\mathbf{g}_{j,k}}\|_{\mathbb{X}} \le \varepsilon_k$. The complexity function $\mathfrak{N}(\mathbf{g}_{j,k})$ is controlled uniformly in j by N_{ε_k}, defined by (2.13). For fixed k the class $\mathfrak{C}_{N_{\varepsilon_k},\mathrm{s}}$ is compact (Proposition 2.9). Therefore, for fixed k, $(G_{\mathbf{g}_{j,k}})_{j\in\mathbb{N}}$ contains a subsequence (again denoted by $(G_{\mathbf{g}_{j,k}})_{j\in\mathbb{N}}$), converging uniformly to some $G_k \in \mathfrak{C}_{N_{\varepsilon_k},\mathrm{s}}$. Now one can take a diagonalization argument, letting k tend to infinity, extracting a convergent subsequence from $(f_j)_{j\in\mathbb{N}}$. □

Appendix B

Proof of Lemma 6.3 As for (a), since

$$\rho_i(t) = \begin{cases} 0, & t < \tau_{i-1}, \\ t - \tau_{i-1}, & t \in J_i, \\ |I|/q, & t > \tau_i. \end{cases} \tag{6.76}$$

(6.19) follows. Regarding (6.20), without loss of generality assume that $t \le t'$ so that

	$t,t' \leqslant \xi_{i-1}$ or $t,t' \geqslant \xi_i$	$t < \xi_{i-1}, t' \in J_i$	$t \leqslant \xi_{i-1}, t' > \xi_i$	$t,t' \in J_i$	$t \in J_i, t' > \xi_i$
$\rho_i(t) - \rho_i(t')$	0	$\xi_{i-1} - t'$	$\lvert J_i\rvert = \frac{\lvert I\rvert}{q}$	$t - t'$	$t - \xi_{i-1} - \lvert I\rvert/q$

Hence $\rho_i(t) \ge \rho_i(t')$, $i = 1, \dots, q$. Specifically, assume that $t' \in J_v$, $t \in J_\ell$. Then

$$\begin{aligned} \sum_{i=1}^{q} |\rho_i(t) - \rho_i(t')| = \sum_{i=\ell}^{v} \rho_i(t) - \rho_i(t') &= |J_\ell| - (t - \tau_{\ell-1}) + |J_{\ell+1}| \\ &\quad + \cdots + |J_{v-1}| + t' - \tau_{v-1} \\ &= t' - t + \tau_v - \tau_\ell - (\tau_{v-1} - \tau_{\ell-1}) = t' - t, \end{aligned}$$

confirming claim (a).

As for (b), we obtain

$$\int_{\underline{t}}^{t} v(s)ds - \sum_{i=1}^{q} \rho_i(t) v_{J_i} = \sum_{i=1}^{q} \int_{J_i} \Big(\chi_{s\le t}(s) - \frac{\rho_i(t)}{|J_i|}\Big) v(s)ds.$$

Now suppose that $t \in J_k$. By (6.76), we have $\Big(\chi_{s\le t}(s) - \frac{\rho_i(t)}{|J_i|}\Big)|_{J_i} = 0$ for $i \le k-1$ while elementary calculations yield

$$\Big| \int_{\tau_{k-1}}^{t} \Big(v(s) - \frac{\rho_k(t) v(s)}{|J_k|}\Big) ds \Big|$$

$$
\begin{aligned}
&= \Big| \frac{\tau_k - t}{\tau_k - \tau_{k-1}} \int_{\tau_{k-1}}^{t} v(s)ds - \frac{t - \tau_{k-1}}{\tau_k - \tau_{k-1}} \int_{t}^{\tau_k} v(s)ds \Big| \\
&\le \frac{\tau_k - t}{\tau_k - \tau_{k-1}} \Big\{ (t - \tau_{k-1}) \|v\|_{L_\infty(J_k)} + (t - \tau_{k-1}) \|v\|_{L_\infty(J_k)} \Big\} \\
&\le 2\|v\|_{L_\infty(J_k)} \frac{(t - \tau_{k-1})(\tau_k - t))}{\tau_k - \tau_{k-1}} \le \frac{|J_k| \|v\|_{L_\infty(J_k)}}{2},
\end{aligned}
$$

which is (6.21).

Concerning (c), let $i(t) := \operatorname{argmin}_{i=1,\dots,q} |t - \xi_i|$. By assumption $\int_{J_i} |v(s) - v(\xi_i)|ds \le |J_i|^2 L'/2$ so that

$$
\begin{aligned}
\Big| \int_0^t v(s)ds - \sum_{i=1}^{q} \rho_i(t) v(\xi_i) \Big| &\le \sum_{i=1}^{i(t)-1} \int_{J_i} |v(s) - v(\xi_i)|ds + \int_{\tau_{i(t)-1}}^{t} |v(s) - v(\xi_{i(t)})|ds \\
&\le i(t) \frac{|J_i|^2 L'}{2} = \frac{i(t)}{q} \frac{|I|^2 L'}{2q}.
\end{aligned}
$$

□

Proof of Lemma 6.6 We consider first the case of fixed x. Using (6.27) and (6.32), one finds for $t \in J_\ell$

$$
\begin{aligned}
&|A_{x,I,q,N}(t, \bar{z}; y) - A_{x,I,q,N}(t, \bar{z}'; y')| \\
&\quad \le \sum_{i=1}^{\ell} \rho_i(t) \|\mathbf{a}\| \max\{|y - y'|, |\bar{z}(\xi_i) - \bar{z}'(\xi_i)|\},
\end{aligned} \tag{6.77}
$$

where $\|\mathbf{a}\|$ plays the role of L. To see the pattern,

$$
\begin{aligned}
&|A^2_{x,I,q,N}(t, \bar{z}; y) - A^2_{x,I,q,N}(t, \bar{z}'; y')| \\
&\quad \le \sum_{i_1=1}^{\ell} \rho_i(t) \|\mathbf{a}\| |A_{x,I,q,N}(\bar{z}(\xi_{i_1}), \xi_{i_1}; y) - A_{x,I,q,N}(\bar{z}(\xi_{i_1}), \xi_{i_1}; y')| \\
&\quad \le \sum_{i_1=1}^{\ell} \rho_{i_1}(t) \|\mathbf{a}\| \Big\{ \sum_{i_2=1}^{i_1} \rho_{i_2}(t) \|\mathbf{a}\| \max\{|y - y'|, |\bar{z}(\xi_{i_2}) - \bar{z}'(\xi_{i_2})|\} \Big\}.
\end{aligned}
$$

Inductively it follows that for $t \in J_\ell$ and $k \in \mathbb{N}$

$$
\begin{aligned}
&|A^k_{x,I,q,N}(t, \bar{z}; y) - A^k_{x,I,q,N}(t, \bar{z}'; y')| \\
&\quad \le \|\mathbf{a}\|^k \sum_{\ell \ge i_1 \ge i_2 \ge \cdots \ge i_k \ge 1} \rho_{i_1}(t) \cdots \rho_{i_k}(t) \max\{|y - y'|, |\bar{z}(\xi_{i_k}) - \bar{z}'(\xi_{i_k})\}|.
\end{aligned}
$$

Invoking (6.22), yields

$$\sum_{\ell\geq i_1\geq i_2\geq\cdots\geq i_k\geq 1}\rho_{i_1}(t)\cdots\rho_{i_k}(t)\leq |I|^k\sum_{\ell\geq i_1\geq i_2\geq\cdots\geq i_k\geq 1}\frac{i_1\cdots i_k}{q^k}\leq\frac{|I|^k}{k!},$$

providing

$$|A^k_{x,I,q,N}(t,\bar z;y)-A^k_{x,I,q,N}(t,\bar z';y')|\leq\frac{(\|\mathbf{a}\||I|)^k}{k!}\max\{|y-y'|,\|\bar z-\bar z'\|_{L_\infty(I;\mathbb{R}^m)}\}.$$

Similarly, for $\bar z=\bar z_x$, $\bar z'=\bar z_{x'}$ (see (6.6)), we obtain

$$\begin{aligned}&|A_{x,I,q,N}(t,\bar z;y)-A_{x',I,q,N}(t,\bar z';y')|\\&\quad\leq|x-x'|+\sum_{i=1}^{\ell}\rho_i(t)\|\mathbf{a}\|\max\{|y-y'|,\|\bar z-\bar z'\|_{L_\infty(I;\mathbb{R}^m)}\},\end{aligned}$$

and hence inductively

$$|A^k_{x,I,q,N}(t,\bar z;y)-A^k_{x',I,q,N}(t,\bar z';y')|\leq\max\{|y-y'|,|x-x'|\}\sum_{\nu=0}^{k}\frac{(\|\mathbf{a}\||I|)^\nu}{\nu!}.$$

Finally, by (6.27) and (6.20), one has for any $t,t'\in I$

$$\begin{aligned}|A_{x,I,q,N}(t,\bar z;y)-A_{x,I,q,N}(t',\bar z;y)|&\leq\sum_{i=1}^{q}|\rho_i(t)-\rho_i(t')||\tilde A_{N,i}(\bar z;y)|\\&\leq\|\mathbf{a}\||t-t'|,\end{aligned}$$

where we have used (6.20) and the definition of $\|\mathbf{a}\|$. This confirms (6.33). The remaining claim follows from Remark 6.4, (6.30). □

Funding Open access funding provided by the Carolinas Consortium.

References

1. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inf. Theory **39**(3), 930–945 (1993)

2. Barron, A., Cohen, A., Dahmen, W., DeVore, R.: Approximation and learning by greedy algorithms. Ann. Stat. **3**(1), 64–94 (2008)
3. Beck, C., Becker, S., Grohs, P., Jaafari, N., Jentzen, A.: Solving the Kolmogorov PDE by means of deep learning. J. Sci. Comput. **88**(3), 1–28
4. Cohen, A., Dahmen, W., DeVore, R., Nichols, J.: Reduced basis greedy selection using random training sets. ESAIM: M2AN **54**(5), 1509–1524 (2020). https://doi.org/10.1051/m2an/2020004
5. Cohen, A., DeVore, R., Petrova, G., Wojtaszczyk, P.: Optimal stable nonlinear approximation Found. Comput. Math. **22**, 607–648 (2020)
6. Dahmen, W., DeVore, R., Grasedyck, L., Süli, E.: Tensor sparsity of solutions to high-dimensional elliptic partial differential equations. Found. Comput. Math. **16**(4), 813–874 (2016). https://doi.org/10.1007/s10208-015-9265-9
7. Daubechies, I., DeVore, R., Foucart, S., Hanin, B., Petrova, G.: Nonlinear approximation and (deep) ReLU networks. Constr. Approx. **55**, 127–172 (2022)
8. Cohen, A., DeVore, R.: Approximation of high-dimensional parametric PDEs. Acta Numer. **24**, 1–159 (2015)
9. DeVore, R.: Nonlinear approximation. Acta Numerica, 1–150 (1998)
10. DeVore, R., Hanin, B., Petrova, G.: Neural network approximation. Acta Numer. **30**, 327–444 (2021). https://doi.org/10.1017/S0962492921000052
11. DeVore, R., Howard, R., Micchelli, C.: Optimal non-linear approximation. Manuscripta Math. **4**, 469–478 (1989)
12. E, W., Ma, C., Wu, L.: The Barron space and the flow-induced function spaces for neural network models. Const. Approx. **55**, 369–406 (2022)
13. Girosi, F., Poggio, T.: Representation properties of networks: Kolmogorov's Theorem is irrelevant. iKun Neural Comput. **1**(4), 465–469 (1989). https://doi.org/10.1162/neco.1989.1.4.465
14. Gribonval, R., Kutyniok, G., Nielsen, M., Voigtlaender, F.: Approximation spaces of deep neural networks. Constr. Approx. **55**, 259–367 (2022)
15. Grohs, P., Herrmann, L.: Deep neural network approximation for high-dimensional parabolic Hamilton–Jacobi–Bellman equations (2021). arXiv:2103.05744v1 [math.NA]
16. Grohs, P., Hornung, F., Jentzen, A., Von Wurstemberger, P.: A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. Memoirs Am. Math. Soc. https://doi.org/10.48550/arXiv.1809.02362
17. Grohs, P., Perekrestenko, D., Elbrächter, D., Bölcskei, H.: Deep neural network approximation theory. IEEE Trans. Inform. Theory (2019). arXiv:1901.02220
18. Grohs, P., Voigtlaender, F.: Proof of the theory-to-practice gap in deep learning via sampling complexity bounds for neural network approximation spaces. Found. Comput. Math. **24**, 1085–1143 (2024)
19. Gühring, I., Kutyniok, G., Petersen, P.: Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$-norms. Anal. Appl. **18**(5), 803–859 (2020). https://doi.org/10.1142/S0219530519410021
20. Hansen, M., Schwab, C.: Sparse adaptive approximation of high dimensional parametric initial value problems. Vietnam J. Math. **41**, 181–215 (2013). https://doi.org/10.1007/s10013-013-0011-9
21. Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., Anandkumar, A.: Neural operator: learning maps between function spaces with applications to PDEs. J. Mach. Learn. Res. **24**, 1–97 (2023)
22. Kurková, V.: Kolmogorov's Theorem is Relevant. Neural Comput. **3**(4), 617–622 (1991)
23. Kutyniok, G., Petersen, P., Raslan, M., Schneider, R.: A theoretical analysis of deep neural networks and parametric PDEs. Constr. Approx. **55**, 73–125 (2022). https://doi.org/10.48550/arXiv.1904.00377
24. Laakmann, F., Petersen, P.: Efficient approximation of solutions of parametric linear transport equations by ReLU DNNs. Adv. Comput. Math. (2021). https://doi.org/10.1007/s10444-020-09834-7
25. Lions, J.P., Perthame, B., Tadmor, E.: A kinetic formulation of multidimensional scalar conservation laws and related equations. J. Am. Math. Soc. **7**(1), 169–191 (1994)
26. Lorentz, G.G.: Metric entropy, widths, and superpositions of functions. Amer. Math. Monthly **69**(6), 469–485 (1962)
27. Mhaskar, H.N., Poggio, T.: Function approximation by deep networks. Commun. Pure Appl. Anal. **19**(8), 4085–4095 (2020). https://doi.org/10.3934/cpaa.2020181
28. Novak, E., Woźniakowski, H.: Approximation of infinitely differentiable multivariate functions is intractable. J. Complexity **25**, 398–404 (2009)

29. Opschoor, J.A.A., Schwab, C., Zech, J.: Exponential ReLU DNN expression of holomorphic maps in high dimension. Constr. Approx. **55**, 537–582 (2022)
30. Petrova, G., Wojtaszczyk, P.: Limitations on approximation by deep and shallow neural networks. J. Mach. Learn. Res. **24**(353), 1–38 (2023)
31. Petersen, P., Voigtlaender, F.: Optimal approximation of piecewise smooth functions using deep relu neural networks. Neural Netw. **108**, 296–330 (2018)
32. Schmidt-Hieber, J.: Nonparametric regression using deep neural networks with ReLu activation. Ann. Statist. **48**(4), 1875–1897 (2020). https://doi.org/10.1214/19-AOS1875(2020)
33. Lu, J., Shen, Z., Yang, H., Zhang, S.: Deep network approximation for smooth functions. SIAM J. Math. Anal. **53**(5), (2020). arXiv:2001.03040
34. Siegel, J.W., Xu, J.: Approximation rates for neural networks with general activation functions. Neural Netw. **128**, 313–321 (2020). https://doi.org/10.1016/j.neunet.2020.05.019
35. Shen, Z., Yang, H., Zhang, S.: Optimal approximation rate of ReLU networks in terms of width and depth. Journal de Mathématiques Pures et Appliquées **157**, 101–135 (2022). https://doi.org/10.1016/j.matpur.2021.07.009
36. Vasseur, A.: Kinetic semi-discretization of scalar conservation laws and convergence by using averaging lemmas. SIAM J. Numer. Anal. **36**(2), 465–474 (1999)
37. Yarotsky, D.: Error bounds for approximations with deep ReLu networks. Neural Netw. **94**, 103–114 (2017)